Towards A Top-Down and Bottom-up Bidirectional Approach to Joint Information Extraction*

Xiaofeng Yu† Irwin King†§ Michael R. Lyu†

†Computer Science & Engineering Dept.
The Chinese University of Hong Kong
{xfyu,king,lyu}@cse.cuhk.edu.hk

§AT&T Labs Research
201 Mission St. Ste 200, San Francisco, CA
irwin@research.att.com

ABSTRACT

Most high-level information extraction (IE) consists of compound and aggregated subtasks. Such IE problems are generally challenging and they have generated increasing interest recently. We investigate two representative IE tasks: (1) entity identification and relation extraction from Wikipedia, and (2) citation matching, and we formally define joint optimization of information extraction. We propose a joint paradigm integrating three factors – segmentation, relation, and segmentation-relation joint factors, to solve all relevant subtasks simultaneously. This modeling offers a natural formalism for exploiting bidirectional rich dependencies and interactions between relevant subtasks to capture mutual benefits. Since exact parameter estimation is prohibitively intractable, we present a general, highly-coupled learning algorithm based on variational expectation maximization (VEM) to perform parameter estimation approximately in a top-down and bottom-up manner, such that information can flow bidirectionally and mutual benefits from different subtasks can be well exploited. In this algorithm, both segmentation and relation are optimized iteratively and collaboratively using hypotheses from each other. We conducted extensive experiments using two real-world datasets to demonstrate the promise of our approach.

Categories and Subject Descriptors

I.5.1 [**Pattern Recognition**]: Models—*Statistical*; H.2.8 [**Database Management**]: Database Applications—*Data mining*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

General Terms

Algorithms, Experimentation

*This work is supported by two grants from the Research Grants Council of the Hong Kong SAR, China (Project No. CUHK 413210 and Project No. CUHK 415410).

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

Keywords

Factor graphs, top-down and bottom-up bidirectional learning, variational inference, joint information extraction

1. INTRODUCTION

Information extraction (IE) aiming at extracting structured information from text or semi-structured sources plays an important role for a wide variety of applications and it has been investigated for decades. Among such IE tasks, high-level IE problems consisting of compound subtasks have become increasingly popular and they present new challenges to research communities. Typically, two key subtasks are *segmentation* which identifies candidate records (e.g., word segmentation, chunking and entity recognition), and *relation* discovery which discovers certain relations between different records (e.g., entity resolution, relation extraction and social relation mining) [23].

The most common and simplest approach to performing compound IE tasks is to use 1-best or K-best pipeline architecture: components are run independently in some order, and there is no feedback from later components to earlier ones [2, 3]. This approach is feed-forward, which is only top-down or bottom-up integrated, and mutual interactions between different components cannot be exploited. Errors cascade and accumulate, and a once-made error can hardly be corrected in the pipeline. Due to this reason, the end-to-end performance is often hampered and upper-bounded.

Ideally, we would like to advocate *joint* information extraction, which is to solve all relevant subtasks in information extraction jointly, that is, all relevant IE subtasks are optimized simultaneously and decisions of them are made together in a single coherent manner. Joint IE aims to handle multiple hypotheses and uncertainty information and to predict many variables at once such that subtasks can aid each other to boost the performance [6, 12, 29, 24, 15, 10, 26, 25]. This is usually very challenging, and often increases the model complexity. It is typically intractable to run a joint model and they sometimes can hurt the performance, since they increase the number of paths to propagate errors. Due to these difficulties, research on building joint approaches is still in the infancy stage.

Recently, a significant amount of work has shown the feasibility and effectiveness of discriminatively-trained probabilistic graphical models for a variety of IE tasks [18, 20]. 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 it can capture complex dependencies between

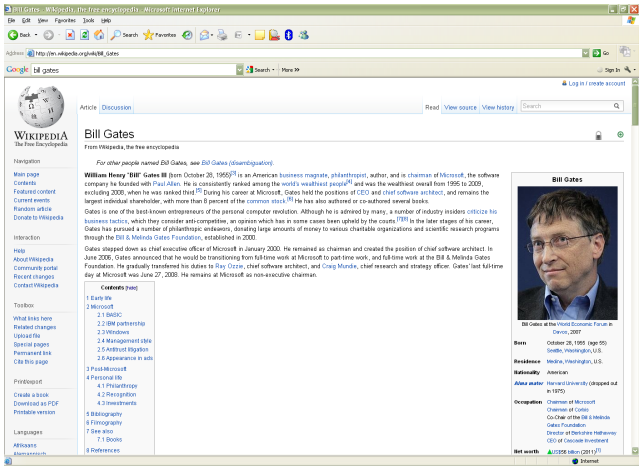

Figure 1: A snapshot of the encyclopedic article about Bill Gates in Wikipedia.

variables. This progress has begun to make the joint learning approach possible. While a number of previous researchers have taken steps toward this direction, there are still various shortcomings: high computational complexity [19], the number of uncertain hypotheses is severely limited [6, 2], the subtasks are only loosely coupled [27], or the approach is feed-forward or top-down integrated and it only allows information to flow in one direction [30].

Exploring bidirectional information and rich interdependencies between relevant subtasks is intuitively appealing. In the following, we use examples from two real-world datasets in our experiments to show the disadvantages of the pipeline architecture and the necessarity of top-down and bottom-up modeling, and to demonstrate the merits of our approach.

1.1 Motivating Examples

1.1.1 Entity Identification and Relation Extraction From Wikipedia

For compound, aggregated IE problems, the availability of robust, flexible, and accurate systems is highly desirable. Wikipedia[1] is the world's largest free online encyclopedia, representing the outcome of a continuous collaborative effort of a large number of volunteer contributors. We investigate the task of identifying entities (e.g., *person*, *location*, and *organization* names) and extracting semantic relationships (e.g., *member_of* and *associate*) between entity pairs in English encyclopedic articles from Wikipedia. For example, Figure 1 gives a snapshot of Wikipedia Web page about the person *Bill Gates*. The basic document is an article, which mainly defines and describes an entity (e.g., *Bill Gates*). This document mentions some other entities (e.g., *Microsoft*, *Paul Allen*, *Seattle*, etc) related to the entity *Bill Gates*. As an illustrative example, consider the following text excerpted from our dataset:

> *George W. Bush was elected President in 2000 as the Republican candidate.*

Clearly, our task consists of two subtasks. First, for entity identification, we need to recognize the entities (both the boundaries and types of them): the *person* name *George W. Bush*, the *year 2000* and the *organization Republican*. Second, for relation extraction, we should extract the *executive* relation between *George W. Bush* and *Republican*. However, the pipeline approach in our experiments cannot extract the *executive* relation between *George W. Bush* and *Republican*, since the *organization* name *Republican* is incorrectly labeled as *miscellaneous* in entity identification stage, and the later relation extraction stage consuming this result also produces erroneous output. From the bottom-up viewpoint, knowing that *Republican* is an *organization* is helpful for the *executive* relation discovery between this entity and the *person George W. Bush*. From the top-down viewpoint, the *executive* relation is a strong evidence indicating a *person* name *George W. Bush* and an *organization Republican*. Modeling top-down and bottom-up simultaneously can therefore explore interdependencies between multiple subtasks, and allow information to flow in both directions to exploit mutual benefits.

1.1.2 Citation Matching

Citation matching requires extracting bibliographic records (e.g., *author*, *title*, and *venue* fields) from citation lists in technical papers (segmentation), and then identifying duplicate records to find the citations referring to the same paper (entity resolution). Citation strings may have different citation styles, different abbreviations, and typographical errors. Correct coreference of a messy citation with a clean citation provides the opportunity for an alignment between these two citations to help the model correctly segment the messy one. Also, coreference between two citations can be assessed more accurately if we can compare well-segmented *title* fields and *venue* fields. Given the following two citations from the Cora database:

> *Hu, Y. & Kibler, D Generation of Attributes for Learning Algorithms, in Proceeding of the 13th National Conference on Artificial Intelligence, p806-811, 1996.*
>
> *Hu, Y., and Kibler, D., Generation of Attributes for Learning Algorithms, Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI96), p.806-811, 1996.*

In the second citation, the *author* field *Hu, Y., and Kibler, D.* and the *title* field *Generation of Attributes for Learning Algorithms* are clearly separated by a comma, and extracting them is fairly straightforward. However, in the first one, there is no clear author-title boundary, and correctly pinpointing it seems very difficult. Large quantities of labeled training data and an extensive lexicon could help, but they are very expensive to obtain, and even available they are far from a guarantee of success. However, if we notice that the two citations are coreferent and the *title* of the second one begins with the substring "*Generation of Attributes for*", we can hypothesize that the *title* of the first one also begins with this substring, allowing us to correctly segment it.

To summarize, what appears to be necessary is a mechanism to consistently integrate top-down and bottom-up processing in a bidirectional manner such that segmentation and entity resolution can aid each other to boost the performance of citation matching. In the following, we summarize our major contributions to show that our approach meets these requirements.

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

1.2 Our Contributions

Inspired by the above motivation and to address the problems such as brittle accumulation of errors in pipeline systems, in this paper we propose a general, strongly-coupled, and bidirectional paradigm, based on conditionally-trained factor graphs for both top-down and bottom-up modeling, to attack the problem of joint information extraction. More specifically, we summarize our major contributions of this paper as follows:

- First, we formally define joint optimization of information extraction. We propose a discriminative framework in which the segmentation-relation joint factor connects bidirectionally with the segmentation and relation factors. This modeling offers a natural formalism for exploiting rich dependencies and interactions between relevant subtasks to capture mutual benefits. It also has several advantages over previous probabilistic graphical models.
- Second, since exact parameter estimation in this model is too expensive, we propose a highly-coupled learning algorithm based on variational expectation maximization (VEM) to perform parameter estimation approximately in a top-down and bottom-up manner, which allows information to flow in both directions and explores mutual benefits from relevant subtasks. Both segmentation and relation are optimized collaboratively to boost the performance. This algorithm is guaranteed to converge in finding the maximum a posteriori (MAP) assignments for model parameters.
- Third, we perform extensive experiments on two important IE tasks, namely, entity identification and relationship extraction from Wikipedia's encyclopedic articles, and citation matching. Our model significantly outperforms state-of-the-art pipeline, integrated and joint models. Some interesting issues, such as the effect of joint factors on performance and efficiency of our approach, are also discussed and analyzed.

2. PROBLEM FORMULATION

Let $\mathbf{x} = \{x_1, x_2, \ldots, x_p\}$ be an observation sequence containing p tokens. Let $\mathbf{s} = \{s_1, s_2, \ldots, s_q\}$ 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 $y_i (y_i \in \mathcal{Y})$ is the label assigned to all tokens of this segment. It is reasonable to assume that segments have positive lengths, and the segment s_i satisfies $0 \leq \alpha_i < \beta_i \leq p$ and $\alpha_{i+1} = \beta_i + 1$. $\mathbf{s}$ essentially models entity candidates, and each segment s_i can be an entity or a non-entity. Without loss of generality, let e_m and e_n $(e_m, e_n \in \mathbf{s})$ be two arbitrary entities in the sequence $\mathbf{x}$, and r_{mn} be the relation assignment between them. $\mathbf{r}$ is the set of relation assignments of all entity pairs in sequence $\mathbf{x}$. $\mathbf{r}$ allows a variety of relations and dependencies, and it is built upon the segmentation $\mathbf{s}$ which models entity candidates. Note that the definitions of $\mathbf{s}$ and $\mathbf{r}$ are general and therefore can be applied to a variety of IE tasks. For example, e_m and e_n can be entity candidates from segments or entire observation sequences. r_{mn} can be a semantic relation (e.g., *employer*) between entity candidates or the boolean coreference variable indicating whether or not two sequences (e.g., paper citations) are referring to each other. r_{mn} can also be an author community or a *friendship* relation in a social Web.

Based on the preliminaries and notations, we define the concepts of segmentation and relation discovery as follows.

DEFINITION 1. *(**Segmentation**). Given an observation sequence $\boldsymbol{x}$, segmentation is the task of assigning segments $\boldsymbol{s}$ to $\boldsymbol{x}$ such that $\boldsymbol{s}^* = \arg\max_{\boldsymbol{s}} P(\boldsymbol{s}|\boldsymbol{x})$.*

DEFINITION 2. *(**Relation Discovery**). For a segmentation $\boldsymbol{s}$ of sequence $\boldsymbol{x}$, relation discovery is the process of extracting and discovering relation $\boldsymbol{r}$ between pairs of entity candidates from $\boldsymbol{s}$ such that $\boldsymbol{r}^* = \arg\max_{\boldsymbol{r}} P(\boldsymbol{r}|\boldsymbol{s}, \boldsymbol{x})$.*

Let $\mathbf{y} = \{\mathbf{r}, \mathbf{s}\}$ be the pair of segmentation $\mathbf{s}$ and relation $\mathbf{r}$. $\mathbf{y}$ must satisfy the condition that both the assignments of the segments and the assignments of the relations of segments are maximized simultaneously. We now formally define the problem of joint information extraction as follows.

DEFINITION 3. *(**Joint Optimization of Information Extraction**). Given an observation sequence $\boldsymbol{x}$, the goal of joint information extraction is to find the assignment $\boldsymbol{y}^* = \{\boldsymbol{r}^*, \boldsymbol{s}^*\}$ that has the maximum a posteriori (MAP) probability*

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

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

Note that this definition is different from pipeline models which perform segmentation and relation in sequential order without capturing interactions between them. This problem is more challenging, and offers new opportunities for information extraction.

3. MODEL

Following the notations in Section 2, we define a joint conditional distribution based on discriminatively-trained factor graphs. Let $\mathcal{G}$ be a factor graph [7] defining a probability distribution over a set of output variables $\mathbf{o}$ conditioned on observation sequences $\mathbf{x}$. $\{\Phi_i\}$ is a set of factors in $\mathcal{G}$, where each factor is defined as the exponential family of an inner product over sufficient statistics $\{f_{ik}(\mathbf{x}_i, \mathbf{o}_i)\}$ and corresponding parameters θ_{ik} as $\Phi_i = \exp\{\sum_k \theta_{ik} f_{ik}(\mathbf{x}_i, \mathbf{o}_i)\}$ [8, 18]. Using parameter tying, the nature of our modeling enables us to partition the factors of $\mathcal{G}$ into three groups, namely the segmentation factor, the relation factor, and the segmentation-relation joint factor. Each factor is a clique template whose parameters are tied. In the following we describe these factors in detail. As we will see, this modeling offers a natural formalism for exploiting top-down and bottom-up bidirectional 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 nonindependent features.

Segmentation factor. The segmentation factor $\phi_S(i, \mathbf{s}, \mathbf{x})$ models segmentations $\mathbf{s}$ in $\mathbf{x}$. We assume that $\phi_S(i, \mathbf{s}, \mathbf{x})$ factorizes according to a set of feature functions $g_k(i, \mathbf{s}, \mathbf{x})$ and a corresponding set of real-valued weights λ_k as

$$\phi_S(i, \mathbf{s}, \mathbf{x}) = \exp\left(\sum_{k=1}^{K} \lambda_k g_k(i, \mathbf{s}, \mathbf{x})\right), \tag{2}$$

where K is the number of feature functions. To effectively capture properties of segmentation, we relax the first-order

Markov assumption to semi-Markov [14] 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, $g_k(i, \mathbf{s}, \mathbf{x}) = g_k(s_{i-1}, s_i, \mathbf{x}) = g_k(y_{i-1}, y_i, \alpha_i, \beta_i, \mathbf{x})$. In addition, transitions within a segment can be non-Markovian.

Relation factor. The relation factor $\phi_R(e_m, e_n, r_{mn})$ models relations r_{mn} $r_{mn} \in \mathbf{r}$ between all possible entity candidate pairs (e_m, e_n), $e_m, e_n \in \mathbf{s}, m \neq n$ in observation sequence $\mathbf{x}$. Similar to the segmentation factor, the relation factor is written as

$$\phi_R(e_m, e_n, r_{mn}) = \exp\left(\sum_{w=1}^{W} \xi_w f_w(e_m, e_n, r_{mn})\right), \quad (3)$$

where W is the number of features, $f_w(e_m, e_n, r_{mn})$ are feature functions, and ξ_w are corresponding weights. The factor $\phi_R(e_m, e_n, r_{mn})$ represents dependencies (e.g., long-distance dependencies, relation transitivity, etc.) between any two entity candidates e_m and e_n. For example, if the same entity is mentioned more than once in an observation sequence, all occurrences probably have the same relation to another entity. Using the relation factor $\phi_R(e_m, e_n, r_{mn})$, evidences for the same entity segments (or entity candidates) to another entity are shared among all their occurrences within the observation sequence.

Segmentation-relation joint factor. Both of segmentation and relation factors are *local*, since they do not take into account dependencies between relevant subtasks. We propose a *global* factor, the segmentation-relation joint factor, to capture both segmentation-to-relation (bottom-up) and relation-to-segmentation (top-down) dependencies. This joint factor $\phi_{\text{SRJ}}(\mathbf{s}, \mathbf{r}, \mathbf{x})$ involves both segmentation and relation hypotheses as its input. It captures the rich and complex interactions between segmentations and relations bidirectionally, which is defined as

$$\phi_{\text{SRJ}}(\mathbf{s}, \mathbf{r}, \mathbf{x}) = \exp\left(\sum_{t=1}^{T} \eta_t q_t(s_{i-1}, s_i, \mathbf{r}, \mathbf{x}) + \sum_{t=1}^{T} \gamma_t h_t(e_m, e_n, r_{mn}, \mathbf{s}, \mathbf{x})\right). \quad (4)$$

The newly introduced feature function $q_t(s_{i-1}, s_i, \mathbf{r}, \mathbf{x})$ exploits relation-to-segmentation (top-down) dependencies, which uses relation hypotheses $\mathbf{r}$ between different segments for segmentation of observation sequence $\mathbf{x}$. Intuitively, knowing the relation between two entity segments is very helpful for segmentation and entity identification. For example, the *employment* relation can only exist between an *organization* and a *person*, and cannot exist between an *organization* and a *location*, or a *location* and a *person*). On the other hand, the function $h_t(e_m, e_n, r_{mn}, \mathbf{s}, \mathbf{x})$ captures segmentation-to-relation (bottom-up) interactions, which uses segmentation information $\mathbf{s}$ for relation discovery. For example, if two segments are labeled as a *location* and a *person*, the semantic relation between them can be *birth_place* or *visited*, but cannot be *employment*. η_t and γ_t are the corresponding real-valued weights for $q_t(\cdot)$ and $h_t(\cdot)$, respectively, and T is the number of features. Notably, this joint factor captures bidirectional interactions and mutual benefits between segmentations and relations. Such dependencies are crucial and modeling them often leads to improved performance.

According to the celebrated Hammersley-Clifford theorem, 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:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\left(\prod_{i=1}^{|\mathbf{s}|} \phi_S(i, \mathbf{s}, \mathbf{x})\right)\left(\prod_{m,n}^{M} \phi_R(e_m, e_m, r_{mn})\right)\left(\prod_{i=1}^{|\mathbf{s}|}\prod_{m,n}^{M} \phi_{\text{SRJ}}(\mathbf{s}, \mathbf{r}, \mathbf{x})\right), \quad (5)$$

where M is the number of arbitrary entity segments in the observation sequence $\mathbf{x}$, $|\mathbf{s}|$ is the number of segments of $\mathbf{x}$, and $Z(\mathbf{x}) = \sum_{\mathbf{y}} \prod_{i=1}^{|\mathbf{s}|} \phi_S(i, \mathbf{s}, \mathbf{x}) \prod_{m,n}^{M} \phi_R(e_m, e_m, r_{mn}) \prod_{i=1}^{|\mathbf{s}|} \prod_{m,n}^{M} \phi_{\text{SRJ}}(\mathbf{s}, \mathbf{r}, \mathbf{x})$ is the normalization factor of our model.

In summary, our model consists of three sub-structures: (1) a semi-Markov chain on the segmentations $\mathbf{s}$ conditioned on the observation sequences $\mathbf{x}$, represented by $\phi_S(i, \mathbf{s}, \mathbf{x})$; (2) potential $\phi_R(e_m, e_n, r_{mn})$ measuring dependencies and relations r_{mn} between two arbitrary entity candidates e_m and e_n from segmentations $\mathbf{s}$; and (3) a fully-connected graph exploiting tight dependencies between segmentations $\mathbf{s}$ and relations $\mathbf{r}$, represented by $\phi_{\text{SRJ}}(\mathbf{s}, \mathbf{r}, \mathbf{x})$. It is particularly notable that our model has a dynamic graphical structure. Since the segments (entity candidates) from the semi-Markov chains are dynamically changed, the structure of relation factor will change correspondingly given different segmentations. Moreover, different structures of the relation factor will also have influence on segmentations. This is different from the conventional semi-CRFs [14].

While some special cases of CRFs are of particular interest, several major elements make our model different. We emphasize on the differences and advantages of our model against others. Most importantly, our model captures bidirectional top-down and bottom-up dependencies between multiple subtasks for joint IE problems. Linear-chain CRFs [8] and semi-CRFs [14] can only perform single IE tasks such as sequence labeling, which lack the ability to capture long-distance dependencies and to represent complex interactions between multiple subtasks. Skip-chain CRFs [17] introduce skip edges to model long-distance dependencies

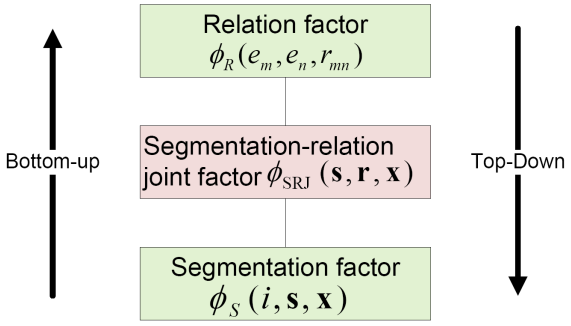

Figure 2: Graphical representation of the proposed model consisting of a bidirectional structure. The segmentation-relation joint factor enables both top-down and bottom-up connections with relation and segmentation factors to explore tight dependencies and mutual benefits for multiple subtasks.

to handle the label consistency problem in single sequence labeling and extraction. 2D CRFs [28] 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 [9] are a class of CRFs with hierarchical tree structure. Our proposed model, on the other hand, has a distinct graphical structure from 2D and hierarchical CRFs. By modeling both segmentations $\mathbf{s}$ and relations $\mathbf{r}$ simultaneously in a single coherent framework, this paradigm offers a natural way for joint information extraction, avoiding the problems such as error propagation occurred in pipeline approaches. Furthermore, this modeling has several advantages over previous probabilistic graphical models, including the employment of semi-Markov chains for efficient segmentation and labeling, the representation of long-range dependencies between different segments, and the capture of rich and complex interactions between relevant subtasks to exploit mutual benefits.

4. TOP-DOWN AND BOTTOM-UP LEARNING

Given independent and identically distributed (i.i.d.) training data $\mathcal{D} = \{\mathbf{x}^i, \mathbf{y}^i\}_{i=1}^{\mathbb{N}}$, where $\mathbf{x}^i$ is the i-th sequence instance, and $\mathbf{y}^i = \{\mathbf{r}^i, \mathbf{s}^i\}$ is the corresponding segmentation and relation assignments. The objective of parameter estimation is to estimate the whole set of model's parameters $\Theta = \{\{\lambda_k\}_{k=1}^K, \{\xi_w\}_{w=1}^W, \{\eta_t\}_{t=1}^T, \{\gamma_t\}_{t=1}^T\}$. Under the i.i.d. assumption, we ignore the summation operator $\sum_{i=1}^N$ in the log-likelihood during the following equations and derivations. We would like to maximize the log-likelihood of the observation given the data:

$$\mathcal{L}(\Theta) = \log P_\Theta(\mathbf{x}) = \log \sum_{\mathbf{s},\mathbf{r}} P_\Theta(\mathbf{s}, \mathbf{r}, \mathbf{x}). \tag{6}$$

The above function does not have a closed-form solution because of the marginalization taking place within the logarithm. Working directly with this function is typically precluded by the need to compute the normalization factor $Z(\mathbf{x})$, which is intractable in our model. We exploit variational approximation methods [5, 20] that offer guarantees in the form of a lower bound on the marginal probabilities. This family of approaches aims to minimize the Kullback-Leibler (KL) divergence between an approximated distribution Q and the target distribution $P_\Theta(\mathbf{s}, \mathbf{r}|\mathbf{x})$ by finding the best distribution Q from some family of distributions for which an inference is feasible. The variational inference method provides a fast, deterministic approximation to otherwise unattainable posteriors. Also its convergence time is independent of dimensionality [20].

Let $Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x})$ be the variational distribution which serves as an approximation of the target distribution $P_\Theta(\mathbf{s}, \mathbf{r}|\mathbf{x})$. According to the mean-field variational theory [5, 21, 4], the optimal solution is the distribution that has the minimum KL divergence between two distributions Q and P. Based on Jensen's inequality we have

$$\begin{aligned}\mathcal{L}(\Theta) &= \log \sum_{\mathbf{s},\mathbf{r}} Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x}) \frac{P_\Theta(\mathbf{s}, \mathbf{r}, \mathbf{x})}{Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x})} \\ &\geq \sum_{\mathbf{s},\mathbf{r}} Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x}) \log \frac{P_\Theta(\mathbf{s}, \mathbf{r}, \mathbf{x})}{Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x})} = KL(Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x}) || P_\Theta(\mathbf{s}, \mathbf{r}, \mathbf{x})).\end{aligned} \tag{7}$$

As can be seen, Equation 7 is the same as maximizing a lower bound on the log-marginal probability $P_\Theta(\mathbf{x})$, with equality when $Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x}) = P_\Theta(\mathbf{s}, \mathbf{r}|\mathbf{x})$. It is easy to obtain the following formulation:

$$\begin{aligned}&\log P_\Theta(\mathbf{x}) - KL(Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x}) || P_\Theta(\mathbf{s}, \mathbf{r}|\mathbf{x})) \\ &\quad = KL(Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x}) || P_\Theta(\mathbf{s}, \mathbf{r}, \mathbf{x})).\end{aligned} \tag{8}$$

According to Equations 7 and 8, optimizing a variational bound on the observed data is equivalent to minimizing the KL divergence between $P_\Theta(\mathbf{s}, \mathbf{r}|\mathbf{x})$ and $Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x})$. This is equivalent to minimizing the KL divergence between the distribution $Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x})$ and the distribution $P_\Theta(\mathbf{s}, \mathbf{r}, \mathbf{x})$. Given the non-negativity property of the KL divergence, the cost function we work is

$$\begin{aligned}\mathscr{L} &= KL(Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x}) || P_\Theta(\mathbf{s}, \mathbf{r}, \mathbf{x})) \\ &= \sum_{\mathbf{s},\mathbf{r}} Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x}) \Big[- \log Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x}) + \log P_\Theta(\mathbf{s}, \mathbf{r}, \mathbf{x}) \Big] \quad (9) \\ &= \mathbb{H}(Q_\nu) + \mathbb{E}_{Q_\nu} \Big\{ \log P_\Theta(\mathbf{s}, \mathbf{r}, \mathbf{x}) \Big\} \quad (10) \\ &\leq \mathcal{L}(\Theta), \quad (11)\end{aligned}$$

where $\mathbb{H}(Q_\nu) = -\sum_{\mathbf{s},\mathbf{r}} Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x}) \log Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x})$ is the entropy of the variational distribution, and $\mathbb{E}_{Q_\nu} \Big\{ \log P_\Theta(\mathbf{s}, \mathbf{r}, \mathbf{x}) \Big\} = \sum_{\mathbf{s},\mathbf{r}} Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x}) \log P_\Theta(\mathbf{s}, \mathbf{r}, \mathbf{x})$ is the expectation with respect to $Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x})$. Notice that the cost $\mathscr{L}$ balances two competing goals: assign values to variables $\mathbf{r}$ and $\mathbf{s}$ with high probability under $P_\Theta(\mathbf{s}, \mathbf{r}, \mathbf{x})$ (the second term), but at the same time be as less committed as possible (the entropy term). Clearly, $\mathscr{L}$ is the lower bound of the log-likelihood $\mathcal{L}(\Theta)$. Thus by maximizing $\mathscr{L}$ we will always recover the log-likelihood of the data $\mathscr{L}^* = \log P_\Theta(\mathbf{x}) - 0$.

For efficient learning, it is critical that the variational family of distributions Q_ν has a tractable form [5, 4]. In the following, we use $Q_\nu(\mathbf{s}, \mathbf{r})$ to denote $Q_\nu(\mathbf{s}, \mathbf{r}|\mathbf{x})$. According to the mean-field variational theory, we assume that $Q_\nu(\mathbf{s}, \mathbf{r})$ forms a factorized distribution; that is, the variables are independent and the joint distribution is a product of single variable marginal probabilities as

$$Q_\nu(\mathbf{s}, \mathbf{r}) = Q_\nu(\mathbf{y}) = \prod_{i \in V_s} Q_{\nu i}(s_i) \prod_{j \in V_r} Q_{\nu j}(r_j), \tag{12}$$

where $\mathbf{s} = \{s_i\}_{i \in V_s}$ and $\mathbf{r} = \{r_j\}_{j \in V_r}$. Let $P_\Theta(\mathbf{s}, \mathbf{r}, \mathbf{x})$ factorize into a product of pairwise potentials depending only on the variables associated with each undirected edge as $P_\Theta(\mathbf{s}, \mathbf{r}, \mathbf{x}) = \prod_{e \in E} \Psi(\mathbf{s}_e, \mathbf{r}_e, \mathbf{x}_e)$. Mean-field inference algorithms exploit this additional factorization structure. Note that we absorb the normalization constant into one of the potentials. The cost function $\mathscr{L}$ reduces to a sum of the following terms as

$$\begin{aligned}\mathscr{L} &= \mathbb{H}(Q_\nu) + \mathbb{E}_{Q_\nu} \Big\{ \log P_\Theta(\mathbf{s}, \mathbf{r}, \mathbf{x}) \Big\} \\ &= \mathbb{H}(Q_\nu) + \mathbb{E}_{Q_\nu} \Big\{ \log \prod_{e \in E} \Psi(\mathbf{s}_e, \mathbf{r}_e, \mathbf{x}_e) \Big\} \quad (13) \\ &= \sum_{i \in V_s} \mathbb{H}(Q_{\nu i}) + \sum_{j \in V_r} \mathbb{H}(Q_{\nu j}) + \sum_{e \in E} \sum_{\mathbf{y}_{e \cap V}} \\ &\quad Q_\nu(\mathbf{y}_{e \cap V}) \log \Psi(\mathbf{s}_e, \mathbf{r}_e, \mathbf{x}_e), \quad (14)\end{aligned}$$

where $Q_\nu(\mathbf{y}_{e \cap V})$ is the variational marginal probability over variables $\mathbf{y} = \{\mathbf{r}, \mathbf{s}\}$ associated with edge e and $V_s \cup V_r = V$.

To optimize the function $\mathscr{L}$, let $\mathbb{E}_{Q_\nu}\{\cdot|y_k\}$ $k \in V$ be the conditional expectation with respect to Q_ν. We provide a

more explicit illustration and the feasibility of evaluating $\mathbb{E}_{Q_\nu}\{\cdot|y_k\}$ in the updates as

$$\mathbb{E}_{Q_\nu}\{\log P_\Theta(\mathbf{s},\mathbf{r},\mathbf{x})|y_k\}$$
$$= \sum_{\{y_i\}_{i\in V\backslash k}} \Big[\prod_{i\in V\backslash k} Q_{\nu i}(y_i)\Big] \log P_\Theta(\mathbf{s},\mathbf{r},\mathbf{x}) \quad (15)$$
$$= \sum_{e\in E} \sum_{\mathbf{y}_{e\cap\{V\backslash k\}}} \sum_{\mathbf{y}_{V\backslash k}} Q_\nu(\mathbf{y}_{e\cap\{V\backslash k\}}) \log \Psi(\mathbf{s}_e,\mathbf{r}_e,\mathbf{x}_e), \quad (16)$$

where $V\backslash k$ is the set of variables other than k and $Q_{\nu i}(y_i) = Q_{\nu i}(s_i)Q_{\nu i}(r_i)$. Note that the expectation specifically does not depend on the variational marginal $Q_{\nu k}(\cdot)$ over y_k. The result is a function of the conditional variable y_k.

To update the k^{th} variational marginal, we view $\mathscr{L}$ as a function of $Q_{\nu k}(\cdot)$ while keeping other marginals fixed. We treat the entropy terms $\mathbb{H}(Q_\nu)$ corresponding to remaining marginals as fixed and appeal to the linearity of expectation $\mathbb{E}_{Q_\nu}\{\cdot\} = \sum_{y_k} Q_{\nu k}(y_k)\mathbb{E}_{Q_\nu}\{\cdot|y_k\}$ to obtain

$$\mathscr{L} = \mathbb{H}(Q_{\nu k}(y_k)) + \sum_{y_k} Q_{\nu k}(y_k)\mathbb{E}_{Q_\nu}\{\log P_\Theta(\mathbf{s},\mathbf{r},\mathbf{x})|y_k\} + \text{const}, \quad (17)$$

where the dependence of $\mathscr{L}$ on the marginal $Q_{\nu k}(y_k)$ is explicit. It is easy to verify via straightforward calculation that maximizing this cost function with respect to the marginal $Q_{\nu k}(y_k)$ yields the following mean field equations for all k as:

$$Q_{\nu k}(y_k) \leftarrow \frac{1}{Z_k} \exp\Big\{\mathbb{E}_{Q_\nu}\{\log P_\Theta(\mathbf{s},\mathbf{r},\mathbf{x})|y_k\}\Big\} \quad (18)$$
$$Z_k \leftarrow \sum_{y_k} \exp\Big\{\mathbb{E}_{Q_\nu}\{\log P_\Theta(\mathbf{s},\mathbf{r},\mathbf{x})|y_k\}\Big\} \quad (19)$$

Obviously, $Q_{\nu k}(y_k)$ is in the form of the exponential family. This property considerably simplifies the complexity and facilitates the computation. Recall that $Q_\nu(\mathbf{s},\mathbf{r}) = Q_\nu(\mathbf{s})Q_\nu(\mathbf{r}) = \prod_{i\in V_s} Q_{\nu i}(s_i)\prod_{j\in V_r} Q_{\nu j}(r_j)$ (Equation 12), we further assume $Q_\nu(\mathbf{s})$ to be of the form $\frac{1}{Z_{Q_\nu(\mathbf{s})}}\prod_{i\in V_s}\phi_i(\mu_i)$ and $Q_\nu(\mathbf{r})$ to be $\frac{1}{Z_{Q_\nu(\mathbf{r})}}\prod_{j\in V_r}\phi_j(\omega_j)$. $Z_{Q_\nu(\mathbf{s})}$ and $Z_{Q_\nu(\mathbf{r})}$ are two local normalization factors. Here, we introduce two variational parameters $\mu = \{\mu_i\}_{i\in V_s}$ and $\omega = \{\omega_j\}_{j\in V_r}$, where μ and ω are associated with segmentation and relation, respectively, and we rewrite $\mathscr{L}$ as $\mathscr{L}(\mu,\omega)$.

We propose a bidirectional learning algorithm based on variational expectation maximization (VEM) to optimize the variational parameters μ and ω efficiently in a collaborative manner such that they can benefit from each other. For example, if we have trained segmentation parameter μ, its decision can guide the learning for relation parameter ω. As shown in Algorithm 1, we summarize the whole parameter estimation procedure as follows: in the E-step, we maximize the variational distributions $Q_{\nu k}(y_k)$, which is accomplished by computing $\mathbb{E}_{Q_\nu}\{\log P_\Theta(\mathbf{s},\mathbf{r},\mathbf{x})|y_k\}$ based on Equations 15 and 16, and by updating $Q_{\nu k}(y_k)$ based on Equations 18 and 19. After we calculate the cost function $\mathscr{L}(\mu,\omega)$ based on Equation 17, we perform bottom-up learning to optimize the relation parameter ω using the hypotheses from segmentations. Here we keep the segmentation parameter μ fixed. In the M-step, we perform top-down learning to optimize the segmentation parameter μ using hypotheses from relations while keeping $Q_{\nu k}(y_k)$ fixed. Such iterative optimization allows information to flow bidirectionally to boost both the segmentation and relation performance. This two-step max-max algorithm leads to a monotonically increasing cost function $\mathscr{L}(\mu,\omega)$ and log-likelihood of data. Consequently, it is guaranteed to converge to an equilibrium state of the KL divergence between Q and P among all distributions Q of the given form $Q_\nu(\mathbf{s}) = \frac{1}{Z_{Q_\nu(\mathbf{s})}}\prod_{i\in V_s}\phi_i(\mu_i)$ and $Q_\nu(\mathbf{r}) = \frac{1}{Z_{Q_\nu(\mathbf{r})}}\prod_{j\in V_r}\phi_j(\omega_j)$. This shows that the algorithm is theoretically sound and correct. Moreover, the variational formulation remains applicable even when we can no longer handle $\log P_\Theta(\mathbf{s},\mathbf{r},\mathbf{x})$, this is superior to the conventional EM algorithm.

Algorithm 1: The variational expectation maximization (VEM) algorithm for bidirectional top-down and bottom-up learning

Input: A set of pairwise potentials $\Psi(\mathbf{s}_e,\mathbf{r}_e,\mathbf{x}_e)$ defining $\log P_\Theta(\mathbf{s},\mathbf{r},\mathbf{x})$, initial potentials for $Q_\nu(\mathbf{s})$ and $Q_\nu(\mathbf{r})$.
Output: Optimized variational parameters μ^* and ω^* for segmentation $\mathbf{s}$ and relation $\mathbf{r}$.
while *equilibrium states or a threshold number of iterations are not reached* **do**
  E-step:
  repeat
    Compute $\mathbb{E}_{Q_\nu}\{\log P_\Theta(\mathbf{s},\mathbf{r},\mathbf{x})|y_k\}$ based on Equation 15 and 16,
    Update $Q_{\nu k}(y_k)$ based on Equation 18 and 19,
    Compute $\mathscr{L}(\mu,\omega)$ based on Equation 17.
    `// Bottom-up learning`
    $\omega^{i+1} = \arg\max_\omega \mathscr{L}(\mu^i,\omega)$.
  until *converge*;
  M-step:
  repeat
    `// Top-down learning`
    $\mu^{i+1} = \arg\max_\mu \mathscr{L}(\mu,\omega^i)$.
  until *converge*;
end
return μ^* *and* ω^*

4.1 Complexity Analysis

We now investigate and analyze the computational complexity of Algorithm 1. Suppose $|V|$ is the number of segmentation and relation variables, and d is the number of distinct values each variable (either $\mathbf{s}$ or $\mathbf{r}$) may take. In Equation 17, the evaluation of the first summation term $\mathbb{H}(Q_{\nu k}(y_k))$ takes $O(|V|d)$. For computing $\mathbb{E}_{Q_\nu}\{\log P_\Theta(\mathbf{s},\mathbf{r},\mathbf{x})|y_k\}$ in the second term of Equation 17, $e\cap\{V\backslash k\}$ is either an empty set or a single node associated with edge e, where each expectation involves at most two variables and there are $|E|$ edges, thus the complexity will be at most $O(|E|d^2)$. For Algorithm 1, suppose the iteration number is $\mathbb{I}$, then the overall computational complexity is $O((|V|d+|E|d^2)\mathbb{I})$.

4.2 Inference

Ideally, the objective of inference is to find the most likely segmentation assignment $\mathbf{s}^*$ and the corresponding most likely relation assignment $\mathbf{r}^*$, that is, to find $\mathbf{y}^* = \arg\max_{\mathbf{y}} P(\mathbf{s},\mathbf{r}|\mathbf{x})$ such that both of them are optimized simultaneously. Unfortunately, exact inference to this problem is generally intractable, since the search space is the Cartesian product

Table 1: Comparative performance of our model, the CRF+CRF, Single MLN, and DCRF models for entity identification from Wikipedia.

Entities	CRF+CRF			Single MLN			DCRF			Our model		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
person	75.33	83.22	79.08	75.94	83.93	79.74	75.96	83.82	79.70	82.91	84.26	**83.58**
location	77.03	69.45	73.04	77.42	70.13	73.59	77.68	70.13	73.71	82.95	81.44	**82.19**
organization	53.78	47.76	50.59	54.11	47.06	50.34	54.55	46.98	50.48	72.43	63.69	**67.78**
date	98.54	97.53	**98.03**	97.79	95.68	96.72	97.98	95.22	96.58	98.90	96.24	97.55
year	97.14	99.10	98.11	98.06	99.12	98.59	98.12	99.09	**98.60**	97.36	99.55	98.44
time	60.00	20.33	30.37	50.00	15.38	23.53	50.00	25.33	33.63	100.0	33.00	**49.62**
number	98.88	60.33	74.94	100.0	66.07	79.57	100.0	70.00	**82.35**	100.0	65.52	79.17
miscellaneous	77.42	80.56	78.96	79.81	84.14	**81.92**	79.81	83.14	81.44	81.50	82.16	81.83
Overall	89.55	88.70	89.12	90.45	90.45	90.45	90.98	90.37	90.67	94.06	93.95	**93.97**

of all possible segmentation and relation assignments. Consequently, approximate inference becomes an inevitable alternative. At the equilibrium state of Algorithm 1, variational distributions $Q_\nu(\mathbf{s})$ and $Q_\nu(\mathbf{r})$ are obtained such that $Q_\nu(\mathbf{s},\mathbf{r}) = Q_\nu(\mathbf{s})Q_\nu(\mathbf{r})$ is an equilibrium state of the KL divergence $KL(Q||P)$. Such kind of inference is straightforward, since the maximum a posterior (MAP) segmentation assignment $\mathbf{s}$ is constructed from the optimized variational parameter μ^*, and the MAP or most likely relation assignment $\mathbf{r}$ is found from the variational parameter ω^*.

5. EXPERIMENTS

5.1 Entity Identification and Relation Extraction

5.1.1 Data and Methodology

Our dataset consists of 1,127 paragraphs from 441 pages from the online encyclopedic articles in Wikipedia. The labeled 7,740 entities are classified into 8 categories, yielding 1,243 *person*, 1,085 *location*, 875 *organization*, 641 *date*, 1,495 *year*, 38 *time*, 59 *number*, and 2,304 *miscellaneous* names. This dataset also contains 4,701 relation instances and 53 labeled relation types, and 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*. The 8 entity categories and 53 relation types are label sets for entity identification and relation extraction in our model. All experiments were performed on the Linux platform, with a 3.2GHz Pentium 4 CPU and 4 GB of memory.

Accurate entities enable features that are naturally expected to be useful to boost relation extraction. A wide range of rich, overlapping features can be exploited in our model. These features include contextual features, part-of-speech tags, morphological features, entity-level dictionary features, and clue word features. Feature conjunctions are also used. In leveraging relation extraction to improve entity identification, we employ a combination of syntactic, entity, keyword, semantic, and Wikipedia characteristic features. More importantly, our model introduces joint factors to capture both top-down and bottom-up dependencies, and function $q_t(\cdot)$ uses relation hypotheses and $h_t(\cdot)$ uses segmentation hypotheses as features. These features capture deep dependencies between entities and relations, and they are natural and effective in enhancing the performance.

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 $F_{\beta=1}$ ($F_{\beta=1}$ is the harmonic mean of P and R and $F_{\beta=1} = \frac{2PR}{P+R}$) for both entity identification and relation extraction. We compare our approach with one pipeline model **CRF+CRF**, one integrated model **Single MLN**, and one joint model **DCRF**. **CRF+CRF** employs one linear-chain CRFs [8] for entity recognition, and another linear-chain CRF for relation prediction. **Single MLN** performs joint inference for both subtasks in a single Markov logic network (MLN) [12], which is a highly expressive language for first-order logic and can conduct relational learning between entity pairs. **DCRF** [19] is a factorial CRF applied to jointly solve the two subtasks. All these models exploit standard parameter learning and inference algorithms in our experiments. To avoid over-fitting, penalization techniques on likelihood are also performed.

Table 2: Comparative performance of different models for relation extraction from Wikipedia.

Method	Accuracy	Precision	Recall	F-measure
CRF+CRF	93.72	70.40	57.85	63.51
Single MLN	93.96	68.54	61.75	64.97
DCRF	93.90	69.30	60.22	64.44
Our model	96.92	72.85	64.25	**68.28**

Table 3: Performance comparison with other top-performing systems on relation extraction.

System	Precision	Recall	F-measure
Culotta et al. [1]	75.53	61.69	67.91
Nguyen et al. [11]	29.07	53.86	37.76
Yu et al. [27]	72.80	59.80	65.66
Our model	72.85	64.25	**68.28**

5.1.2 Experimental Results and Discussions

Table 1 shows the performance of entity identification and Table 2 shows the overall performance of relation extraction[2] of different models, respectively. For relation extraction in Table 2, we also recorded the token-wise labeling accuracy. Our model substantially outperforms all baseline models on F-measure for both entity identification and relation extraction, and it is statistically significantly better (p-value < 0.05 with a 95% confidence interval) according to McNemar's paired tests. The pipeline model **CRF+CRF** suffers from pipeline inherent inferiority such as brittle accumulation of errors. For example, this model cannot correctly extract relations between mis-recognized entities. As discussed, it performs entity identification and relation extraction independently without considering the mutual correlations between them, leading to reduced performance. By modeling interactions between two subtasks, boosted performance can be achieved, as illustrated by the integrated model **Single MLN** and the joint model **DCRF**. The **Single MLN** model captures dependencies between entities and relations via first-order logic; however, limitations of first-order logic

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

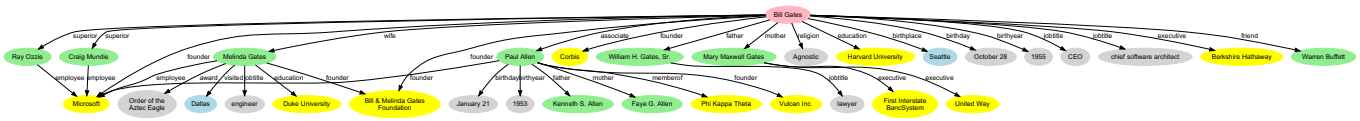

Figure 3: An example of identified entities and extracted semantic relationships from Wikipedia's encyclopedic article about *Bill Gates* by using our model. The major person name *Bill Gates* is in pink and other person names are in green color. Organizations are in yellow and locations are in blue, and other types of entities are in gray. Semantic relations between entities are also labeled.

make it difficult to specify a relation factor that utilizes the uncertain output of segmentation [15, 22]. Joint inference in **Single MLN** is only weakly coupled and does not enforce transitivity, since the logic formulas only examine pairs of consecutive labels, not whole fields. As can be seen, our model achieves stronger interactions between two subtasks, which is strongly coupled and bidirectional. The **DCRF** model applies loopy belief propagation (LBP) for approximate learning and inference, which is inherently unstable and may cause convergence problems. Consequently, training a **DCRF** model with unobserved nodes (hidden variables) makes this approach difficult to optimize. Figure 3 illustrates an example of identified entities and extracted semantic relationships from Wikipedia's encyclopedic article about *Bill Gates* by employing our model. As can be seen, different entity types are in different colors, and the relations between them are also linked and labeled. Interestingly, these results are versatile for a variety of applications, such as Web data mining, social network analysis and mining, etc.

A large number of engineered systems were developed for identifying relations of interest. Table 3 compares our results with some recently published results on the same dataset. Notably, our approach outperforms previous ones given that we deal with a fairly more challenging problem involving both entity identification and relation extraction. Similar to [13], these systems assume that the golden-standard entities are already known or extracted from text without errors, and they only perform relation extraction (due to this reason, we only compare the performance on relation extraction.). Unfortunately, such assumption is not valid in practice. As a result, our model is more applicable to real-world IE tasks.

5.1.3 Effect of Joint Factors on Performance

We investigate the nature and effectiveness of segmentation-relation joint factors and Figure 4 demonstrates their feasibility in our modeling. It shows that the joint factors consistently enhance precision, recall, and F-measure for both entity identification and relation extraction subtasks. For example, the joint factors significantly improve the overall F-measure by 2.65% for entity identification. Our approach demonstrates its merits by using joint factors to explore bidirectional tight interactions between segmentations and relations and by optimizing them collaboratively in a top-down and bottom-up manner, resulting in improved performance.

5.1.4 Efficiency

Table 4 summarizes the efficiency of different models. The pipeline **CRF+CRF** takes the least time for learning, due to its simple pipeline architecture. Compared to **Single MLN**, the running time of our model is only increased slightly, which is reasonable to apply to real IE problems. It is particularly notable that our model takes much less time than the joint model **DCRF**. Specifically, our model is over an order of magnitude (approximately 17.5 times) faster than **DCRF** for running. When the graph has large tree-width as in our case, the LBP algorithm in **DCRF** is inefficient and slow to converge.

Table 4: Efficiency comparison of different models on learning time (min.) and inference time (min.).

Method	Learning time	Inference time
CRF+CRF	15.30	0.20
Single MLN	35.95	2.67
DCRF	677.67	0.33
Our model	39.83	0.30

5.2 Citation Matching

5.2.1 Data and Methodology

We apply the Cora dataset to evaluate our approach. This dataset contains 1295 citations and 134 clusters (sets of citations that refer to the same paper), and each citation has three fields – *author*, *title*, and *venue*. We run three-fold cross-validation on this dataset. Segmentation is evaluated by P, R, and F_1. For entity resolution, we measure both pairwise P, R, F_1 and cluster recall. Cluster recall is the fraction of clusters that are correctly output by the system after taking transitive closure from pairwise decisions.

A wide range of rich features can be exploited in our model. For segmentation, these features largely consider field-level similarity using a number of string and token-based comparison metrics (e.g., string edit distance, tf-idf over tokens and n-grams, etc.). We also include feature conjunctions, specialized features for *author* and *title* matching, and global features based on distance metrics for entire citations. In leveraging coreference to improve segmentation, we employ a combination of local (e.g., contextual and morphological), layout, lexicon membership features.

For performance comparison, the **CRF+CRF** model applies first CRF for segmentation, and another CRF for entity resolution (which views resolution as a pairwise classification problem). For the **Single MLN** model, we follow [12] to design it, engaging features mentioned above. Moreover, we also compare the performance of our model with some recently published results on the same dataset.

5.2.2 Experimental Results and Discussions

Our comparative results are shown in Table 5 and Table 6, demonstrating the promise of our approach with significant improvements on both segmentation and coreference compared with the three baseline models **CRF+CRF**, **Single MLN**, **DCRF**, and other previously published results.

Table 5 shows the improvements on F-measure for segmentation, where we list both the overall performance and

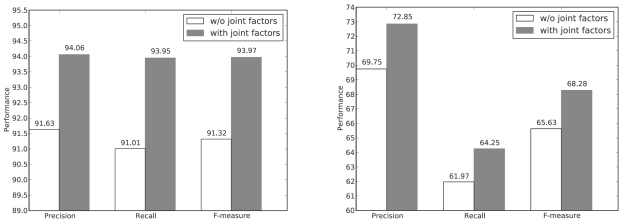

Figure 4: Performance comparison of joint factors on entity identification (left) and relation extraction (right) from Wikipedia.

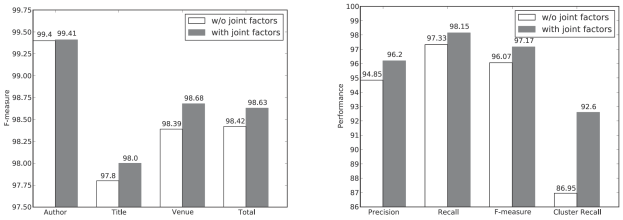

Figure 5: Performance comparison of joint factors on segmentation (left) and entity resolution (right) in citation matching.

Table 5: Comparative performance of different models for segmentation in citation matching.

Method	Author	Title	Venue	Total
Isolated MLN [12]	99.30	97.30	98.20	98.20
Single MLN [12]	99.50	97.60	98.30	98.40
CRF+CRF	98.77	97.02	97.56	97.66
Single MLN	99.39	97.79	98.36	98.41
DCRF	99.40	97.82	98.45	98.47
Our model	99.41	98.00	98.68	**98.63**

Table 6: Comparative performance of different models for entity resolution in citation matching.

Method	P	R	F_1	Cluster Recall
Fellegi-Sunter [16]	78.00	97.70	86.70	62.70
Single MLN [12]	94.30	97.00	95.60	78.10
CRF+CRF	93.10	94.65	93.87	76.32
Single MLN	94.84	97.22	96.02	85.15
DCRF	94.92	97.69	96.29	85.77
Our model	96.20	98.15	**97.17**	**92.60**

the performance on the three fields. Our model outperforms earlier results such as **Isolated MLN** [12] and **Single MLN** [12]. Compared to the three baseline models, the relative error reduction (RER) is 41.45%, 13.84% and 11.68%, respectively. Note that the difference between our **Single MLN** model and the one in [12] is that we engage different features. Table 6 compares the performance of entity resolution for different models on both metrics of F_1 and cluster recall. Our model, which concurrently solves the citation matching task, easily outperforms previously published results in [16] and [12]. It also outperforms the three baseline models by 3.30%, 1.15%, and 0.88% in pairwise F_1. Even though the **Single MLN** model in [12] captures interactions between segmentation and coreference, it is only a weak interaction. Since the logic formulae in [12] only examine pairs of consecutive labels but not whole fields – failing to utilize information from predicted field range and non-consecutive words in the field. As can be seen, our model is highly-coupled and achieves stronger interaction between multiple subtasks. Importantly, Table 6 shows that our approach allows cluster recall to improve substantially, resulting in an improvement of up to 6.83% compared to **DCRF** model. This is particularly notable given that cluster recall is more strict than the pairwise F_1 metric.

5.2.3 *Effect of Joint Factors on Performance*

Figure 5 illustrates the benefits of the joint factors in our modeling for citation matching. This figure demonstrates the bidirectionality of joint factors using segmentation to aid coreference and vice versa, which is highly coupled and information can flow in both directions to capture mutual benefits and strong interactions between segmentation and resolution. Compared to the pairwise F-measure, the cluster recall is boosted substantially (up to 5.65%) by joint factors. This is particularly interesting as it shows that exploiting joint factors is much more accurate under the strict metric.

5.2.4 *Efficiency*

Table 7 lists the running times (for both training and inference) for **CRF+CRF**, **Single MLN**, **DCRF**, and our model (we cannot compare it with the models in [16] and [12] since experimental settings are different). The running time of our model is reasonably slower than that of the pipeline model **CRF+CRF**, and comparable to that of **Single MLN**, illustrating the efficiency of our approach. However, the **DCRF** model takes 458.83 minutes to converge, which is very slow. This disadvantage limits the ability of **DCRF** for real-world IE problems to a large extent.

6. RELATED WORK

Some work has been dedicated to improving the pipeline architecture [2, 3]. Finkel et al.[2] modeled pipelines as Bayesian networks, with each low level task corresponding to a variable in the network. This architecture has the drawback that it only allows information to flow in one direction. Hollingshead and Roark [3] proposed pipeline iteration, using output from later stages of a pipeline to constrain earlier stages of the same pipeline, but it lacks the ability to model internal tight dependencies between stages. However, all these approaches suffer from inherent problems such as brittle accumulation of errors caused by their pipeline architecture.

Integrated and joint models exploring mutual benefits on different subtasks have shown great promise, where several closely related approaches have been proposed. Recently, Zhu et al. [30] proposed an integrated probabilistic approach to Web page understanding. Nevertheless, this model is feed-forward or top-down integrated and it only allows information to flow in one direction. Ko et al. [6] proposed a joint answer ranking framework based on probabilistic graphical models for question answering. However, they employed N-best list for inference procedure which is a restricted approximation for the full distribution of large-output components. Thus, the number of uncertain hypotheses in their framework is severely limited. Yu et al.[27] integrated two sub-models, semi-CRFs and MLNs, together, but they are only loosely coupled in that the parameter estimation is performed separately and the inference information can only flow in one direction, which is similar to [2]. Luo et al. [10]

Table 7: Efficiency comparison of different models on learning time (min.) and inference time (min.).

Method	Learning time	Inference time
CRF+CRF	10.33	0.17
Single MLN	23.80	1.85
DCRF	458.83	0.30
Our model	26.67	0.25

combined Web classification and Web IE based on the CRF model. However, since it was defined according to the DOM tree structure for Web pages, this model cannot be applied to more general tasks that we are investigating.

Furthermore, Poon and Domingos [12] performed joint inference in a single MLN to citation matching, and Sutton et al.[19] proposed dynamic CRFs to jointly solve part-of-speech tagging and NP chunking tasks. As shown in our experiments, limitations of first-order logic make the model in [12] only loosely coupled. On the other hand, the dynamic CRF model in [19] includes complex graphical structure and high computational complexity, which may cause convergence problems. Our proposed model is highly coupled and bidirectional, and considerably outperforms both of them.

7. CONCLUSION AND FUTURE WORK

We presented a strongly-coupled, bidirectional approach to the problem of joint information extraction. We introduced joint factors to capture top-down and bottom-up bidirectional tight correlations and dependencies between subtasks, and we proposed a learning algorithm based on VEM to perform parameter estimation approximately in a top-down and bottom-up manner. This algorithm allows information to flow in both directions and explores mutual benefits from multiple subtasks. Experimental results on two real-world datasets exhibit that our model significantly outperforms recent state-of-the-art pipeline, integrated and joint models while also running much faster than the joint models. Several interesting issues, such as the effect of joint factors on performance and the efficiency of our approach are analyzed and discussed as well. This approach allows extensive further investigation, both for parameter learning and inference. We also plan to apply and test our model to other real-world IE applications.

8. REFERENCES

[1] A. Culotta, A. McCallum, and J. 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.

[2] J. R. Finkel, C. D. Manning, and A. 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.

[3] K. Hollingshead and B. Roark. Pipeline iteration. In *Proceedings of ACL-07*, pages 952–959, Prague, Czech Republic, 2007.

[4] T. S. Jaakkola. Tutorial on variational approximation methods. In *Advanced Mean Field Methods: Theory and Practice*, pages 129–159. MIT Press, 2000.

[5] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical methods. *Machine Learning*, 37:183–233, 1999.

[6] J. Ko, L. Si, and E. Nyberg. A probabilistic graphical model for joint answer ranking in question answering. In *Proceedings of SIGIR-07*, pages 343–350, Amsterdam, The Netherlands, 2007.

[7] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[8] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML-01*, pages 282–289, 2001.

[9] L. Liao, D. Fox, and H. Kautz. Extracting places and activities from GPS traces using hierarchical conditional random fields. *International Journal of Robotics Research*, 26:119–134, 2007.

[10] P. Luo, F. Lin, Y. Xiong, Y. Zhao, and Z. Shi. Towards combining Web classification and Web information extraction: a case study. In *Proceedings of KDD-09*, pages 1235–1244, Paris, France, 2009.

[11] D. P. T. Nguyen, Y. Matsuo, and M. Ishizuka. Relation extraction from Wikipedia using subtree mining. In *Proceedings of AAAI-07*, pages 1414–1420, Vancouver, British Columbia, Canada, 2007.

[12] H. Poon and P. Domingos. Joint inference in information extraction. In *Proceedings of AAAI-07*, pages 913–918, Vancouver, British Columbia, Canada, 2007.

[13] F. Reichartz, H. Korte, and G. Paass. Semantic relation extraction with kernels over typed dependency trees. In *Proceedings of KDD-10*, pages 773–782, New York, 2010.

[14] S. Sarawagi and W. W. Cohen. Semi-Markov conditional random fields for information extraction. In *Proceedings of NIPS-04*, 2004.

[15] S. Singh, K. Schultz, and A. Mccallum. Bi-directional joint inference for entity resolution and segmentation using imperatively-defined factor graphs. In *Proceedings of ECML/PKDD-09*, pages 414–429, Bled, Slovenia, 2009.

[16] P. Singla and P. Domingos. Entity resolution with Markov logic. In *Proceedings of ICDM-06*, pages 572–582, 2006.

[17] C. Sutton and A. 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.

[18] C. Sutton and A. McCallum. An introduction to conditional random fields for relational learning. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press, 2006.

[19] C. Sutton, A. McCallum, and K. Rohanimanesh. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *Journal of Machine Learning Research*, 8:693–723, 2007.

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

[21] W. Wiegerinck. Variational approximations between mean field theory and the junction tree algorithm. In *Proceedings of UAI-2000*, pages 626–633, San Francisco, CA, 2000.

[22] K. Yoshikawa, S. Riedel, M. Asahara, and Y. Matsumoto. Jointly identifying temporal relations with Markov logic. In *Proceedings of ACL-09*, pages 405–413, Suntec, Singapore, 2009.

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

[24] X. Yu and W. Lam. An integrated probabilistic and logic approach to encyclopedia relation extraction with multiple features. In *Proceedings of COLING-08*, pages 1065–1072, Manchester, United Kingdom, 2008.

[25] X. Yu and W. Lam. Bidirectional integration of pipeline models. In *Proceedings of AAAI-10*, pages 1045–1050, Atlanta, Georgia, USA, 2010.

[26] X. Yu and W. Lam. Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach. In *Proceedings of COLING-10*, pages 1399–1407, Beijing, China, 2010.

[27] X. Yu, W. Lam, and B. Chen. An integrated discriminative probabilistic approach to information extraction. In *Proceedings of CIKM-09*, pages 325–334, Hong Kong, China, 2009.

[28] J. Zhu, Z. Nie, J.-R. Wen, B. Zhang, and W.-Y. Ma. 2D conditional random fields for Web information extraction. In *Proceedings of ICML-05*, pages 1044–1051, Bonn, Germany, 2005.

[29] J. Zhu, Z. Nie, B. Zhang, and J.-R. Wen. Dynamic hierarchical Markov random fields for integrated Web data extraction. *Journal of Machine Learning Research*, 9:1583–1614, 2008.

[30] J. Zhu, B. Zhang, Z. Nie, J.-R. Wen, and H.-W. Hon. Webpage understanding: an integrated approach. In *Proceedings of KDD-07*, pages 903–912, San Jose, California, USA, 2007.