A Framework Based on Graphical Models with Logic for Chinese Named Entity Recognition *

Xiaofeng YU **Wai LAM** **Shing-Kit CHAN**

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

Abstract

Chinese named entity recognition (NER) has recently been viewed as a *classification* or *sequence labeling* problem, and many approaches have been proposed. However, they tend to address this problem *without* considering linguistic information in Chinese NEs. We propose a new framework based on probabilistic graphical models with first-order logic for Chinese NER. First, we use Conditional Random Fields (CRFs), a standard and theoretically well-founded machine learning method based on undirected graphical models as a base system. Second, we introduce various types of domain knowledge into Markov Logic Networks (MLNs), an effective combination of first-order logic and probabilistic graphical models for validation and error correction of entities. Experimental results show that our framework of probabilistic graphical models with first-order logic significantly outperforms the state-of-the-art models for solving this task.

1 Introduction

Named entity recognition (NER) is the task of identifying and classifying phrases that denote certain types of named entities (NEs), such as person names (PERs), locations (LOCs) and organizations (ORGs) in text documents. It is a well-established task in the NLP and data mining communities and is regarded as crucial technology for many higher-level applications, such as information extraction, question answering, information retrieval and knowledge management. The NER problem has generated much interest and great progress has been made, as evidenced by its inclusion as an understanding task to be evaluated in the Message Understanding Conference (MUC), the Multilingual Entity Task (MET) evaluations, and the Conference on Computational Natural Language Learning (CoNLL).

Compared to European-language NER, Chinese NER seems to be more difficult (Yu *et al.*, 2006). Recent approaches to Chinese NER are a shift away from manually constructed rules or finite state patterns towards machine learning or statistical methods. However, rule-based NER systems lack robustness and portability. Statistical methods often suffer from the problem of data sparsity, and machine learning approaches (e.g., Hidden Markov Models (HMMs) (Bikel *et al.*, 1999; Zhou and Su, 2002), Support Vector Machines (SVMs) (Isozaki and Kazawa, 2002), Maximum Entropy (MaxEnt) (Borthwick, 1999; Chieu and Ng, 2003), Transformation-based Learning (TBL) (Brill, 1995) or variants of them) might be unsatisfactory to learn linguistic information in Chinese NEs. Current state-of-the-art models often view Chinese NER as a *classification* or *sequence labeling* problem *without* considering the linguistic and structural information in Chinese NEs. They assume that entities are independent, however in most cases this assumption does not hold because of the existing relationships among the entities. They seek to locate and identify named entities in text by sequentially classifying tokens (words or characters) as to whether or not they participate in an NE, which is sometimes prone to noise and errors.

In fact, Chinese NEs have distinct linguistic characteristics in their composition and human beings usually use prior knowledge to recognize NEs. For example, about 365 of the highest frequently used surnames cover 99% Chinese surnames (Sun *et al.*, 1995). Some LOCs contain location salient words, while some ORGs contain organization salient words. For the LOC "香港特区/Hong Kong Special Region", "香港/Hong Kong" is the name part and "特区/Special Region" is the salient word. For the ORG "香港特区政府/Hong Kong Special Region Government", "香港/Hong Kong" is the LOC name part, "特区/Special Region" is the LOC salient word and "政府/Government" is the ORG salient word. Some ORGs contain one or more PERs, LOCs and ORGs. A more complex exam-

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 Nos: CUHK 4179/03E and CUHK4193/04E) and the Direct Grant of the Faculty of Engineering, CUHK (Project Codes: 2050363 and 2050391). This work is also affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

ple is the nested ORG "北京市海淀区清华大学计算机学院/School of Computer Science, Tsinghua University, Haidian District, Beijing City" which contains two ORGs "清华大学/Tsinghua University" and "计算机学院/School of Computer Science" and two LOCs "北京市/Beijing City" and "海淀区/Haidian District". The two ORGs contain ORG salient words "大学/University" and "学院/School", while the two LOCs contain LOC salient words "市/City" and "区/District" respectively.

Inspired by the above observation, we propose a new framework based on probabilistic graphical models with first-order logic which treats Chinese NER [1] as a *statistical relational learning* (SRL) problem and makes use of domain knowledge. First, we employ Conditional Random Fields (CRFs), a discriminatively trained undirected graphical model which has theoretical justification and has been shown to be an effective approach to segmenting and labeling sequence data, as our base system. We then exploit a variety of domain knowledge into Markov Logic Networks (MLNs), a powerful combination of logic and probability, to validate and correct errors made in the base system. We show how a variety of domain knowledge can be formulated as first-order logic and incorporated into MLNs. We use three Markov chain Monte Carlo (MCMC) algorithms, including Gibbs sampling, Simulated Tempering, as well as MC-SAT, and Maximum a posteriori/Most Probable Explanation (MAP/MPE) algorithm for probabilistic inference in MLNs. Experimental results show that our framework based on graphical models with logic yields substantially better NER results, leading to a relative error reduction of up to 23.75% on the F-measure over state-of-the-art models. McNemar's tests confirm that the improvements we obtained are statistically highly significant.

2 State of the Art

2.1 CRF Model for Chinese NER

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. CRFs have the great flexibility to encode a wide variety of arbitrary, non-independent features and to straightforwardly combine rich domain knowledge. Furthermore, they are discriminatively trained, and are often more accurate than generative models, even with the same features. CRFs have been successfully applied to a number of real-world tasks, including NP chunking (Sha and Pereira, 2003), Chinese word segmentation (Peng *et al.*, 2004), information extraction (Pinto *et al.*, 2003; Peng and McCallum, 2004), named entity identification (McCallum and Li, 2003; Settles, 2004), and many others.

[1] In this paper we only focus on PERs, LOCs and ORGs. Since temporal, numerical and monetary phrases can be well identified with rule-based approaches.

Recently, CRFs have been shown to perform exceptionally well on Chinese NER shared task on the third SIGHAN Chinese language processing bakeoff (SIGHAN-06) (Zhou *et al.*, 2006; Chen *et al.*, 2006b,a). We follow the state-of-the-art CRF models using features that have been shown to be very effective in Chinese NER, namely the current character and its part-of-speech (POS) tag, several characters surrounding (both before and after) the current character and their POS tags, current word and several words surrounding the current word.

We also observe some important issues that significantly influence the performance as follows:

Window size: The primitive window size we use is 5 (2 characters preceding the current character and 2 following the current character). We extend the window size to 7 but find that it slightly hurts. The reason is that CRFs can deal with non-independent features. A larger window size may introduce noisy and irrelevant features.

Feature representation: For character features, we use character identities. For word features, BIES representation (each character is beginning of a word, inside of a word, end of a word, or a single word) is employed.

Labeling scheme: The labeling scheme can be BIO, BIOE or BIOES representation. In BIO representation, each character is tagged as either the beginning of a named entity (B), a character inside a named entity (I), or a character outside a named entity (O). In BIOE, the last character in an entity is labeled as E while in BIOES, single-character entities are labeled as S. In general, BIOES representation is more informative and yields better results than both BIO and BIOE.

2.2 Error Analysis

Even though the CRF model is able to accommodate a large number of well-engineered features which can be easily obtained across languages, some NEs, especially LOCs and ORGs are difficult to identify due to the lack of linguistic or structural characteristics. Since predictions are made token by token, some typical and serious tagging errors are still made, as shown below:

- **ORG is incorrectly tagged as LOC**: In Chinese, many ORGs contain location information. The CRF model only tags the location information (in the ORGs) as LOCs. For example, "唐山理工学院/Tangshan Technical Institute" and "海南省省委/Hainan Provincial Committee " are ORGs and they contain LOCs "唐山/Tangshan" and "海南省/Hainan Province", respectively. "唐山/Tangshan" and "海南省/Hainan Province" are only incorrectly tagged as LOCs. This affects the tagging performance of both ORGs and LOCs.

- **LOC is incorrectly tagged as ORG**: The LOCs "悉尼歌剧院/Sydney Opera" and "北京体育馆/Beijing Gymnasium" are mistakenly tagged as ORGs by the CRF model without taking into account the location salient words "歌剧院/Opera" and "体育馆/Gymnasium".

- **The boundary of entity is tagged incorrectly**: This mistake occurs for all the entities. For example, the PER "汤姆·克鲁斯/Tom Cruise" may be tagged as a PER "汤姆/Tom"; the LOC "不来梅/Bremen" may be tagged as a LOC "来梅/Laimei", which is a meaningless word; the ORG "华为公司/Huawei Corporation" may be tagged as an ORG "华为/Huawei". The reasons for these errors are both complicated and varied. However, some of them are related to linguistic knowledge.

- **Common nouns are incorrectly tagged as entities**: For example, the two common nouns "现代数学/Modern Mathematics" and "格兰士微波炉/Galanz Microwave Oven" may be improperly tagged as a LOC and an ORG. Some tagging errors could be easily rectified. Take the erroneous ORG "市委组织，/City Committee Organizes," for example, intuitively it is not an ORG since an entity cannot span any punctuation.

3 Our Proposed Framework

3.1 Overview

We propose a framework based on probabilistic graphical models with first-order logic for Chinese NER. As shown in Figure 1, the framework is composed of three main components. The CRF model is used as a base model. Then we incorporate domain knowledge that can be well formulated into first-order logic to extract entity candidates from CRF results. Finally, the Markov Logic Network (MLN), an undirected graphical model for *statistical relational learning*, is used to validate and correct the errors made in the base model. We begin by briefly reviewing the necessary background of MLNs, including weight learning and inference.

3.2 Markov Logic Networks

A Markov Network (also known as Markov Random Field) is a model for the joint distribution of a set of variables (Pearl, 1988). It is composed of an undirected graph $G = (V, E)$ and a set of real-valued potential functions ϕ_k. A First-Order Knowledge Base (KB) (Genesereth and Nislsson, 1987) is a set of sentences or formulas in first-order logic.

A Markov Logic Network (MLN) (Richardson and Domingos, 2006) is a KB with a weight attached to each formula (or clause). Together with a set of constants representing objects in the domain, it species a ground Markov Network containing one feature for each possible grounding of a first-order formula F_i in the KB, with the corresponding weight w_i. The basic idea in MLNs is that: when a world violates one formula in the KB it is less probable, but not impossible. The fewer formulas a world violates, the more probable it is. The weights associated with the formulas in an MLN jointly determine the probabilities of those formulas (and vice versa) via a *log-linear model*. An MLN is a statistical relational model that defines a probability distribution over Herbrand interpretations (possible worlds), and can

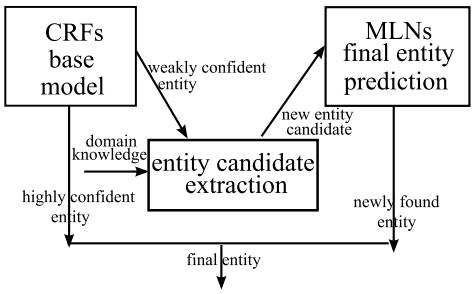

Figure 1: Framework Overview

be thought of as a *template* for constructing Markov Networks. Given different sets of constants, it will produce different networks. These networks will have certain regularities in structure and parameter given by the MLN and they are called ground Markov Networks. Suppose `Peter(A)`, `Smith(B)` and `IBM(X)` are 3 constants, a KB and generated features are listed in Table 1. The formula `Employ(x,y)⇒Person(x),Company(y)` means `x` is employed by `y` and `Colleague(x,y)⇒Employ(x,z)∧Employ(y,z)` means `x` and `y` are colleagues if they are employed by the same company. Figure 2 shows the graph of the ground Markov network defined by the formulas in Table 1 and the 3 constants `Peter(A)`, `Smith(B)` and `IBM(X)`. The probability distribution over possible worlds x specified by the ground Markov Network $M_{L,C}$ is given by

$$P(X = x) = \frac{1}{Z}\exp(\sum w_i n_i(x)) = \frac{1}{Z}\prod \phi_i\left(x_{\{i\}}\right)^{n_i(x)} \quad (1)$$

where $n_i(x)$ is the number of true groundings of F_i in x, $x_{\{i\}}$ is the true value of the atoms appearing in F_i, and $\phi_i\left(x_{\{i\}}\right) = e^{w_i}$.

In the case of Chinese NER, a named entity can be connected to another named entity for instance, because they share the same location salient word. Thus in an undirected graph, two node types exist, the LOC nodes and the location salient word nodes. The links (edges) indicate the relation (LOCs contain location salient words) between them. This representation can be well expressed by MLNs.

However, one problem concerning relational data is, how to extract useful relations for Chinese NER. There are many kinds of relations between NEs, some relations are critical to the NER problem while others not. Another problem that we address is whether these relations can be formulated in first-order logic and combined in MLNs. In Section 3.3, we exploit domain knowledge. We will show how these knowledge can capture essential characteristics of Chinese NEs and can be well and concisely formulated in first-order logic in Section 3.4.

Table 1: Example of a KB and Generated Features

Fist-Order Logic (KB)	Generated Features
∀ x,y Employ(x,y)⇒Person(x),Company(y)	Employ(Peter,IBM)⇒Person(Peter),Company(IBM) Employ(Smith,IBM)⇒Person(Smith),Company(IBM)
∀ x,y,z Colleague(x,y)⇒ Employ(x,z)∧Employ(y,z)	Colleague(Peter,Smith)⇒ Employ(Peter,IBM) ∧Employ(Smith,IBM)

3.2.1 Learning Weights

Given a relational database, MLN weights can in principle be learned generatively by maximizing the likelihood of this database on the closed world assumption. The gradient of the log-likelihood with respect to the weights is

$$\frac{\partial}{\partial w_i}\log P_w(X=x) = n_i(x) - \sum P_w(X=x')n_i(x') \tag{2}$$

where the sum is over all possible databases x', and $P_w(X = x')$ is $P(X = x')$ computed using the current weight vector $w = (w_1, ..., w_i, ...)$. Unfortunately, computing these expectations can be very expensive. Instead, we can maximize the *pseudo-log-likelihood* of the data more efficiently. If x is a possible database and x_l is the *l*th ground atom's truth value, the *pseudo-log-likelihood* of x given weights w is

$$\log P^*_w(X=x) = \sum_{l=1}^{n} \log P_w(X_{l=x_l} \mid MB_x(X_l)) \tag{3}$$

where $MB_x(X_l)$ is the state of X_l's *Markov blanket* [2] in the data. Computing Equation 3 and its gradient does not require inference over the model, and is therefore much faster. We can optimize the *pseudo-log-likelihood* using the limited-memory BFGS algorithm (Liu and Nocedal, 1989).

3.2.2 Inference

If F_1 and F_2 are two formulas in first-order logic, C is a finite set of constants including any constants that appear in F_1 or F_2, and L is an MLN, then

$$\begin{aligned} P(F_1 \mid F_2, L, C) &= P(F_1 \mid F_2, M_{L,C}) \\ &= \frac{P(F_1 \wedge F_2 \mid M_{L,C})}{P(F_2 \mid M_{L,C})} \\ &= \frac{\sum_{x \in \chi_{F_1} \cap \chi_{F_2}} P(X=x \mid M_{L,C})}{\sum_{x \in \chi_{F_2}} P(X=x \mid M_{L,C})} \end{aligned} \tag{4}$$

where χ_{F_i} is the set of worlds where F_i holds, and $P(x \mid M_{L,C})$ is given by Equation 1. The question of whether a knowledge base entails a formula F in first-order logic is the question of whether $P(F \mid L_{\mathrm{KB}}, C_{\mathrm{KB},F}) = 1$, where L_{KB} is the MLN obtained by assigning infinite weight to all the formulas in KB, and $C_{\mathrm{KB},F}$ is the set of all constants appearing in KB or F.

[2] The Markov blanket of a node is the minimal set of nodes that renders it independent of the remaining network; in a MLN, this is simply the node's neighbors in the graph.

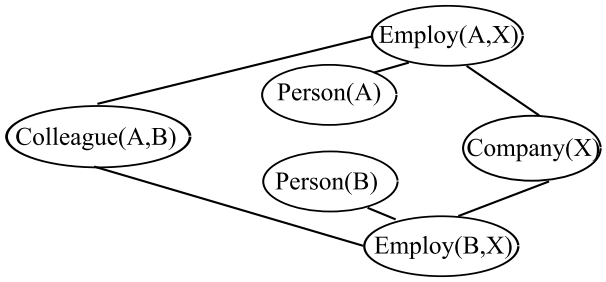

Figure 2: A Ground Markov network defined by the formulas in Table 1 and the constants Peter(A), Smith(B) and IBM(X).

A large number of efficient inference techniques are applicable to MLNs. The most widely used approximate solution to probabilistic inference in MLNs is Markov chain Monte Carlo (MCMC) (Gilks *et al.*, 1996). In this framework, the Gibbs sampling algorithm is to generate an instance from the distribution of each variable in turn, conditional on the current values of the other variables. The key to the Gibbs sampler is that one only considers univariate conditional distributions-the distribution when all of the random variables but one are assigned fixed values. One way to speed up Gibbs sampling is by Simulated Tempering (Marinari and Parisi, 1992), which performs simulation in a *generalized ensemble*, and can rapidly achieve an equilibrium state. Poon and Domingos (2006) proposed MC-SAT, an inference algorithm that combines ideas from MCMC and satisfiability. MC-SAT works well and is guaranteed to be sound, even when deterministic or near-deterministic dependencies are present in real-world reasoning.

Besides MCMC framework, maximum a posteriori (MAP) inference can be carried out using a weighted satisfiability solver like MaxWalkSAT. It is closely related to maximum likelihood (ML), but employs an augmented optimization objective which incorporates a prior distribution over the quantity one wants to estimate. MAP estimation can therefore be seen as a regularization of ML estimation.

3.3 Domain Knowledge

We incorporate various kinds of domain knowledge via MLNs to predict the newly extracted NE candidates from

CRF hypotheses. We extract 165 location salient words and 843 organization salient words from Wikipedia[3] and the LDC Chinese-English bi-directional NE lists compiled from Xinhua News database, as shown in Table 2. We also make a punctuation list which contains 18 items and some stopwords which Chinese NEs cannot contain. The stopwords are mainly conjunctions, auxiliary and functional words. We extract new NE candidates from the CRF results according to the following consideration:

- Definitely, if a chunk (a series of continuous characters) occurs in the training data as a PER or a LOC or an ORG, then this chunk should be a PER or a LOC or an ORG in the testing data. In general, a unique string is defined as a PER, it cannot be a LOC somewhere else.
- Obviously, if a tagged entity ends with a location salient word, it is a LOC. If a tagged entity ends with an organization salient word, it is an ORG.
- If a tagged entity is close to a subsequent location salient word, probably they should be combined together as a LOC. The closer they are, the more likely that they should be combined.
- If a series of consecutive tagged entities are close to a subsequent organization salient word, they should probably be combined together as an ORG because an ORG may contain multiple PERs, LOCs and ORGs.
- Similarly, if there exists a series of consecutive tagged entities and the last one is tagged as an ORG, it is likely that all of them should be combined as an ORG.
- Entity length restriction: all kinds of tagged entities cannot exceed 25 Chinese characters.
- Stopword restriction: intuitively, all tagged entities cannot comprise any stopword.
- Punctuation restriction: in general, all tagged entities cannot span any punctuation.
- Since all NEs are proper nouns, the tagged entities should end with noun words.
- The CRF model tags each token (Chinese character) with a conditional probability. A low probability implies a low-confidence prediction. For a chunk with low conditional probabilities, all the above assumptions are adopted (The marginal probabilities are normalized, and probabilities lower than the user-defined threshold are regarded as low conditional probabilities).

All the above domain knowledge can be formulated as first-order logic to construct the structure of MLNs. And all the extracted chunks are accepted as new NE candidates (or common nouns). We train an MLN to recognize them.

[3]http://en.wikipedia.org/wiki/.

Table 2: Domain Knowledge for Chinese NER

Location Salient Word	Organization Salient Word
自治区/Municipality	百货公司/Department Store
火车站/Railway Station	理工学院/Technical Institute
宾馆/Hotel	旅行社/Travel Agency
公园/Park	出版社/Press
高原/Plateau	人事部/Personnel Department
省/Province	银行/Bank
镇/Town	大学/University
市/City	市委/City Committee
Stopword	**Punctuation**
仍然/still	。
但是/but	？
非常/very	，
的/of	；
等/and so on	：
那/that	！

3.4 First-Order Logic Representation

We declared 14 *predicates* (`person(candidate)`, `location(candidate)`, `organization(candidate)`, `endwith(candidate, salientword)`, `closeto(candidate, salientword)`, `containstopword(candidate)`, `containpunctuation(candidate)`, etc) and specified 15 first-order formulas (See Table 3 for some examples) according to the domain knowledge described in Section 3.3. For example, we used `person(candidate)` to specify whether a candidate is a PER. *Formulas* are recursively constructed from atomic formulas using logical connectives and quantifiers. They are constructed using four types of symbols: *constants*, *variables*, *functions*, and *predicates*. *Constant* symbols represent objects in the domain of interest (e.g., "北京/Beijing" and "上海/Shanghai" are LOCs). *Variable* symbols (e.g., `r` and `p`) range over the objects in the domain. To reduce the size of ground Markov Network, variables and constants are *typed*; for example, the variable `r` may range over candidates, and the constant "北京/Beijing" may represent a LOC. *Function* symbols represent mappings from tuples of objects to objects. *Predicate* symbols represent relations among objects (e.g., `person`) in the domain or attributes of objects (e.g., `endwith`). A *ground atom* is an atomic formula all of whose arguments are ground terms (terms containing no variables). For example, the ground atom `location(北京市)` conveys that "北京市/Beijing City" is a LOC.

For example in Table 3, "乌市/Wu City" is mis-tagged as an ORG by the CRF model, but it contains the location salient word "市/City". So it is extracted as a new entity candidate, and the corresponding formula `endwith(r, p)`$\wedge$`locsalientword(p)`$\Rightarrow$`location(r)` means if `r` ends with a location salient word `p`, then it is a LOC. Besides the formulas listed in Table 3, we also specified logic such as `person(p)`$\Rightarrow$`!(location(p) v organization(p))`, which means a candidate `p` can

Table 3: Examples of NE Candidates and First-Order Formulas

Mis-tagged NEs	New NE Candidates	First-Order Logic
希拉里[common noun]	希拉里	occurperson(p)⇒person(p)
凡尔赛[PER]	凡尔赛	occurlocation(p)⇒location(p)
一汽集团[common noun]	一汽集团	occurorganization(p)⇒organization(p)
乌市[ORG]	乌市	endwith(r,p)∧locsalientword(p)⇒location(r)
英政府[LOC]	英政府	endwith(r,p)∧orgsalientword(p)⇒organization(r)
北海[LOC]花园	北海花园	closeto(r,p)∧locsalientword(p)⇒location(r)
瑞士[LOC]联邦	瑞士联邦	closeto(r,p)∧orgsalientword(p)⇒organization(r)
市区的酒店[LOC]	市区的酒店	containstopword(p)⇒!(person(p) v location(p) v organization(p))
“百帮”服务中心[ORG]	“百帮”服务中心	containpunctuation(p)⇒!(person(p) v location(p) v organization(p))

only belong to one class.

We assume that the relational database contains only binary relations. Each extracted NE candidate is represented by one or more strings appearing as arguments of ground atoms in the database. The goal of NE prediction is to determine whether the candidates are entities and the types of entities (query predicates), given the evidence predicates and other relations that can be deterministically derived from the database. As we will see, despite their simplicity and consistency, these first-order formulas incorporate the essential features for NE prediction.

4 Experiments

4.1 Dataset

We used People's Daily corpus (January-Jun, 1998) in our experiments, which contains approximately 357K sentences, 156K PERs, 219K LOCs and 87K ORGs, respectively. We did some modifications on the original data to make it cleaner. We enriched some tags so that the abbreviation proper nouns are well labeled. We preprocessed some nested names to make them in better form. We also processed some person names. We enriched tags for different kinds of person names (e.g., Chinese and transliterated names) and separated consecutive person names.

4.2 The Baseline NER System

We use CRFs to build a character-based Chinese NER system, with features described in Section 2.1. To avoid overfitting, we penalized the log-likelihood by the commonly used zero-mean Gaussian prior over the parameters. In addition, we exploit clue word features which can capture non-local dependencies. This gives us a competitive baseline CRF model using both local and non-local information for Chinese NER.

For clue word features, we employ 412 career titles (e.g., 总统/President, 教授/Professor, 警察/Police), 59 family titles (e.g., 爸爸/Father, 妹妹/Sister), 33 personal pronouns (e.g., 你们/Your, 我们/We) and 109 direction words (e.g., 以北/North, 南部/South) to represent non-local information. Career titles, family titles and personal pronouns may imply a nearby PER and direction words may indicate a LOC or an ORG. Figure 3 illustrates an example of non-local dependency.

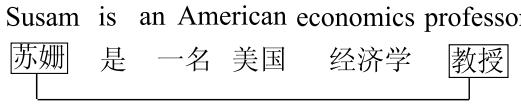

Figure 3: An Example of Non-local Dependency. The Career Title “教授” Indicates a PER “苏姗”

We do not take the advantage of using the golden-standard word segmentation and POS tagging provided in the original corpus, since such information is hardly available in real text. Instead, we use an off-the-shelf Chinese lexical analysis system, the open source ICTCLAS (Zhang *et al.*, 2003), to segment and POS tag the corpus. This module employs a hierarchical Hidden Markov Model (HHMM) and provides word segmentation, POS tagging (labels Chinese words using a set of 39 tags) and unknown word recognition. It performs reasonably well, with segmentation precision recently evaluated at 97.58%. The recall of unknown words using role tagging is over 90%.

We use one-month corpus for training and 9-day corpus for testing. Table 4 shows the experimental results.

4.3 NER System Based on Graphical Models with Logic

To test the effectiveness of our proposed model, we extract all the NEs (19,879 PERs, 25,661 LOCs and 11,590 ORGs) from the training corpus. An MLN training database, which consists of 14 predicates, 16,620 constants and 97,992 ground atoms was built.

The MLNs were trained using a Gaussian prior with zero mean and unit variance on each weight to penalize the pseudo-likelihood, and with the weights initialized at the mode of the prior (zero). During MLN learning, each formula is converted to Conjunctive Normal Form (CNF), and a weight is learned for each of its clauses. The weight

Table 4: Chinese NER by CRF Model

	Precision	Recall	$F_{\beta=1}$
Character features			
PER	92.88%	79.42%	85.62
LOC	90.95%	82.88%	86.73
ORG	88.16%	83.86%	85.96
Overall	90.92%	82.07%	86.27
Character+Word			
PER	93.27%	82.99%	87.83
LOC	91.49%	85.16%	88.21
ORG	88.94%	84.79%	86.82
Overall	91.48%	84.46%	87.83
Character+Word+POS			
PER	92.17%	90.64%	91.40
LOC	90.56%	89.74%	90.15
ORG	89.15%	85.19%	87.12
Overall	90.76%	89.13%	89.94
All features			
PER	92.12%	90.57%	91.34
LOC	90.62%	89.74%	90.18
ORG	89.72%	85.44%	87.53
Overall	90.89%	89.16%	90.02

Table 5: Chinese NER by Graphical Models with Logic

	Precision	Recall	$F_{\beta=1}$	RER
CRF Baseline				
PER	92.12%	90.57%	91.34	
LOC	90.62%	89.74%	90.18	
ORG	89.72%	85.44%	87.53	
Overall	90.89%	89.16%	90.02	
Graphical Models (GS Inference)				
PER	93.52%	93.32%	93.42	
LOC	93.19%	91.91%	92.55	
ORG	90.16%	90.71%	90.43	
Overall	92.70%	92.09%	**92.39**	23.75%
Graphical Models (ST Inference)				
PER	93.52%	93.32%	93.42	
LOC	93.19%	91.91%	92.55	
ORG	90.16%	90.71%	90.43	
Overall	92.70%	92.09%	**92.39**	23.75%
Graphical Models (MC-SAT Inference)				
PER	93.52%	93.32%	93.42	
LOC	93.19%	91.91%	92.55	
ORG	90.16%	90.71%	90.43	
Overall	92.70%	92.09%	**92.39**	23.75%
Graphical Models (MAP/MPE Inference)				
PER	92.87%	93.15%	93.01	
LOC	93.15%	91.61%	92.37	
ORG	90.56%	89.10%	89.82	
Overall	92.57%	91.58%	**92.07**	20.54%

of a clause is used as the mean of a Gaussian prior for the learned weight. These weights reflect how often the clauses are actually observed in the training data.

We extract 529 entity candidates to construct the MLN testing database, which contains 2,543 entries and these entries are used as evidence for inference. Inference is performed by grounding the minimal subset of the network required for answering the query predicates. We employed 3 MCMC algorithms: Gibbs sampling (GS), Simulated Tempering (ST) as well as MC-SAT, and the MAP/MPE algorithm for inference and the comparative NER results are shown. The probabilistic graphical models greatly outperform the CRF model stand-alone by a large margin. It can be seen from Table 5, the probabilistic graphical models integrating first-order logic improve the precision and recall for all kinds of entities, thus boosting the overall F-measure. We achieve a 23.75% relative error reduction (RER) on F-measure by using 3 MCMC algorithms and a 20.54% RER by using MAP/MPE algorithm, over an already competitive CRF baseline. We obtained the same results using GS, ST and MC-SAT algorithms. MCMC algorithms yields slightly better results than the MAP/MPE algorithm.

4.4 Significance Test

Ideally, comparisons among NER systems would control for feature sets, data preparation, training and test procedures, parameter tuning, and estimate the statistical significance of performance differences. Unfortunately, reported results sometimes leave out details needed for accurate comparisons.

We give statistical significance estimates using McNemar's paired tests [4] (Gillick and Cox, 1989) on labeling disagreements for CRF model and graphical probabilistic models that we evaluated directly.

Table 6 summarizes the correctness of the labeling decisions between the models with a 95% confidence interval (CI). These tests suggest that the graphical probabilistic models are significantly more accurate and confirm that the gains we obtained are statistically highly significant.

Table 6: McNemar's Tests on Labeling Disagreements

Null Hypothesis	95% CI	p-value
Proposed Model (GS) vs. CRFs	5.71-9.52	$< 1 \cdot 10^{-6}$
Proposed Model (ST) vs. CRFs	5.71-9.52	$< 1 \cdot 10^{-6}$
Proposed Model (MC-SAT) vs. CRFs	5.71-9.52	$< 1 \cdot 10^{-6}$
Proposed Model (MAP/MPE) vs. CRFs	4.50-7.37	$< 1 \cdot 10^{-6}$

5 Related Work

As a well-established task, Chinese NER has been studied extensively and a number of techniques for this task have been reported in the literature. Most recently, the trend in Chinese NER is to use improved machine learning approaches, or to integrate various kinds of useful evidences, features, or resources.

Fu and Luke (2005) presented a lexicalized HMM-based approach to unifying unknown word identification

[4] Most researchers refer to statistically significant as $p < 0.05$ and statistically highly significant as $p < 0.001$.

and NER as a single tagging task on a sequence of known words. Although lexicalized HMMs was shown to be superior to standard HMMs, this approach has some disadvantages: it is a purely statistical model and it suffers from the problem of data sparseness. And the model fails to tag some complicated NEs (e.g., nested ORGs) correctly due to lack of domain adaptive techniques. The F-measures of LOCs and ORGs are only 87.13 and 83.60, which show that there is still a room for improving.

A method of incorporating heuristic human knowledge into a statistical model was proposed in (Wu *et al.*, 2005). Here Chinese NER was regarded as a probabilistic tagging problem and the heuristic human knowledge was used to reduce the searching space. However, this method assumes that POS tags are golden-standard in the training data and heuristic human knowledge is often ad hoc. These drawbacks make the method unstable and highly sensitive to POS errors; and when golden-standard POS tags are not available (this is often the case), it may degrade the performance.

Cohen and Sarawagi (2004) proposed a semi-Markov model which combines a Markovian, HMM-like extraction process and a dictionary component. This process is based on sequentially classifying segments of several adjacent words. However, this technique requires that entire segments have the same class label, while our technique does not. Moreover, compared to a large-scale dictionary, our domain knowledge is much easier to obtain.

However, all the above models treat NER as classification or sequence labeling problem. To the best of our knowledge, MLNs have not been previously used for NER problem. To our knowledge, we first view Chinese NER as a *statistical relational learning* problem and exploit domain knowledge which can be concisely formulated in MLNs, allowing the training and inference algorithms to be directly applied to them.

6 Conclusion and Future Work

The contribution of this paper is three-fold. First, we formulate Chinese NER as a *statistical relational learning* problem and propose a new framework incorporating probabilistic graphical models and first-order logic for Chinese NER which achieves state-of-the-art performance. Second, We incorporate domain knowledge to capture the essential features of the NER task via MLNs, a unified framework for SRL which produces a set of weighted first-order clauses to predict new NE candidates. To the best of our knowledge, this is the first attempt at using MLNs for the NER problem in the NLP community. Third, our proposed framework can be extendable to language-independent NER, due to the simplicity of the domain knowledge we could access. Directions for future work include learning the structure of MLNs automatically and using MLNs for information extraction (e.g., entity relation extraction).

References

Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. An algorithm that learns what's in a name. *Machine Learning*, 34(1-3):211–231, February 1999.

Andrew Borthwick. *A Maximum Entropy Approach to Named Entity Recognition*. PhD thesis, New York University, September 1999.

Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565, 1995.

Aitao Chen, Fuchun Peng, Roy Shan, and Gordon Sun. Chinese named entity recognition with conditional probabilistic models. In *5th SIGHAN Workshop on Chinese Language Processing*, Australia, July 2006.

Wenliang Chen, Yujie Zhang, and Hitoshi Isahara. Chinese named entity recognition with conditional random fields. In *5th SIGHAN Workshop on Chinese Language Processing*, Australia, July 2006.

Hai Leong Chieu and Hwee Tou Ng. Named entity recognition with a maximum entropy approach. In *Proceedings of CoNLL-03*, 2003.

William W. Cohen and Sunita Sarawagi. Exploiting dictionaries in named entity extraction: Combining semi-Markov extraction processes and data integration methods. In *Proceedings of ACM-SIGKDD 2004*, 2004.

Guohong Fu and Kang-Kwong Luke. Chinese named entity recognition using lexicalized HMMs. *ACM SIGKDD Explorations Newsletter*, 7:19–25, June 2005.

Michael R. Genesereth and Nils J. Nislsson. *Logical foundations of artificial intelligence*. Morgan Kaufmann Publishers Inc., San Mateo, CA, 1987.

W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman and Hall, London, UK, 1996.

L. Gillick and Stephen Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Proceedings of ICASSP-89*, pages 532–535, 1989.

Hideki Isozaki and Hideto Kazawa. Efficient support vector classifiers for named entity recognition. In *Proceedings of COLING-02*, pages 1–7, Taipei, Taiwan, 2002.

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. Morgan Kaufmann, San Francisco, CA, 2001.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989.

Enzo Marinari and Giorgio Parisi. Simulated Tempering: A new Monte Carlo scheme. *Europhysics Letters*, 19:451–458, 1992.

Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of CoNLL-03*, 2003.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, 1988.

Fuchun Peng and Andrew McCallum. Accurate information extraction from research papers using conditional random fields. In *Proceedings of HLT-NAACL 2004*, pages 329–336, 2004.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of COLING-04*, pages 562–568, 2004.

David Pinto, Andrew McCallum, Xing Wei, and W. Bruce Croft. Table extraction using conditional random fields. In *Proceedings of ACM SIGIR-03*, 2003.

Hoifung Poon and Pedro Domingos. Sound and efficient inference with probabilistic and deterministic dependencies. In *Proceedings of AAAI-06*, Boston, Massachusetts, July 2006. The AAAI Press.

Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.

Burr Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, Geneva, Switzerland, 2004.

Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. In *Proceedings of HLT-NAACL 2003*, pages 213–220, 2003.

Maosong Sun, Changning Huang, Haiyan Gao, and Jie Fang. Identifying Chinese names in unrestricted texts. *Journal of Chinese Information Processing*, 1995.

Youzheng Wu, Jun Zhao, Bo Xu, and Hao Yu. Chinese named entity recognition based on multiple features. In *Proceedings of HLT-EMNLP 2005*, 2005.

Xiaofeng Yu, Marine Carpuat, and Dekai Wu. Boosting for Chinese named entity recognition. In *5th SIGHAN Workshop on Chinese Language Processing*, Australia, July 2006.

Hua Ping Zhang, Qun Liu, Xue-Qi Cheng, Hao Zhang, and Hong Kui Yu. Chinese lexical analysis using Hierarchical Hidden Markov Model. In *2nd SIGHAN Workshop on Chinese Language Processing*, volume 17, pages 63–70, 2003.

Guodong Zhou and Jian Su. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of ACL-02*, pages 473–480, Philadelphia, USA, 2002.

Junsheng Zhou, Liang He, Xinyu Dai, and Jiajun Chen. Chinese named entity recognition with a multi-phase model. In *5th SIGHAN Workshop on Chinese Language Processing*, Australia, July 2006.