Chinese NER Using CRFs and Logic for the Fourth SIGHAN Bakeoff [*]

Xiaofeng YU Wai LAM Shing-Kit CHAN Yiu Kei WU Bo CHEN

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

Abstract

We report a high-performance Chinese NER system that incorporates Conditional Random Fields (CRFs) and first-order logic for the fourth SIGHAN Chinese language processing bakeoff (SIGHAN-6). Using current state-of-the-art CRFs along with a set of well-engineered features for Chinese NER as the base model, we consider distinct linguistic characteristics in Chinese named entities by introducing 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. Our submitted results achieved consistently high performance, including the first place on the CityU open track and fourth place on the MSRA open track respectively, which show both the attractiveness and effectiveness of our proposed model.

1 Introduction

We participated in the Chinese named entity recognition (NER) task for the fourth SIGHAN Chinese language processing bakeoff (SIGHAN-6). We submitted results for the open track of the NER task. Our official results achieved consistently high performance, including the first place on the CityU open track and fourth place on the MSRA open track. This paper presents an overview of our system due to space limit. A more detailed description of our model is presented in (Yu *et al.*, 2008).

Our Chinese NER system combines the strength of two graphical discriminative models, Conditional Random Fields (CRFs) and Markov Logic Networks (MLNs). First, we employ CRFs, a discriminatively trained undirected graphical model which has been shown to be an effective approach to segmenting and labeling sequence data, as our base system. Second, we model the linguistic and structural information in Chinese named entity composition. We exploit a variety of domain knowledge which can capture essential characteristics of Chinese named entities into Markov Logic Networks (MLNs), a powerful combination of first-order logic and probability, to (1) validate and correct errors made in the base system and (2) find and extract new entity candidates. These domain knowledge is easy to obtain and can be well and concisely formulated in first-order logic and incorporated into MLNs.

2 Conditional Random Fields as Base Model

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 been shown to perform well on Chinese NER shared task on SIGHAN-4 (Zhou *et al.* (2006), Chen *et al.* (2006a), Chen *et al.* (2006b)). We employ CRFs as the base model in our framework. In this base model, we design features similar to the state-of-the-art CRF models for Chinese NER. We use character features, word segmentation features, part-of-speech (POS) features, and dictionary features, as described below.

Character features: These features are the current character, 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.

Word segmentation and POS features: We train our own model for conducting Chinese word segmentation and POS tagging. We employ a unified framework to integrate cascaded Chinese word segmentation and POS tagging tasks by joint decoding that guards against vi-

[*] 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, CUHK4193/04E, and CUHK4128/07) 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.

olations of those hard-constraints imposed by segmentation task based on dual-layer CRFs introduced by Shi and Wang (2007).

We separately train the Chinese word segmentation and POS tagging CRF models using 8-month and 2-month PKU 2000 corpus, respectively. The original PKU 2000 corpus contains more than 100 different POS tags. To reduce the training time for POS tagging experiment, we merge some similar tags and obtain only 42 tags finally. For example, {`ia`, `ib`, `id`, `in`, `iv`}→`i`. We use the same features as described in (Shi and Wang, 2007), except that we do not use the HowNet features for word segmentation. Instead, we use max-matching segmentation features based on a word dictionary. This dictionary contains 445456 words which are extracted from People's Daily corpus (January-June, 1998), CityU, MSRA, and PKU word segmentation training corpora in SIGHAN-6. For decoding, we first perform individual decoding for each task. We then set 10-best segmentation and POS tagging results for reranking and joint decoding in order to find the most probable joint decodings for both tasks.

Dictionary features: We obtain a named entity dictionary extracted from People's Daily 1998 corpus and PKU 2000 corpus, which contains 68305 PERs, 28408 LOCs and 55596 ORGs. We use the max-matching algorithm to search whether a string exists in this dictionary.

In summary, we list the features used for our CRF base model in Table 1. Besides the unigram feature template, CRFs also allow bigram feature template. With this template, a combination of the current output token and previous output token (bigram) is automatically generated.

We use CRF++ toolkit (version 0.48) (Kudo, 2005) in our experiments. We find that setting the cut-off threshold f for the features not only decreases the training time, but improves the NER performance. CRFs can use the features that occurs no less than f times in the given training data. We set $f = 5$ in our system.

We extend the **BIO** representation for the chunk tag which was employed in the CoNLL-2002 and CoNLL-2003 evaluations. We use the **BIOES** representation in which each character is tagged as either the beginning of a named entity (**B** tag), a character inside a named entity (**I** tag), the last character in an entity (**E** tag), single-character entities (**S** tag), or a character outside a named entity (**O** tag). We find that **BIOES** representation is more informative and yields better results than **BIO** representation.

3 Markov Logic Networks as Error Correction Model

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.

Table 1: Feature template for CRF model.

Character features	(1.1) $C_n, n \in [-2, 2]$ (1.2) $C_nC_{n+1}, n \in [-2, 1]$
Word features	(1.3) $W_n, n \in [-3, 3]$ (1.4) $W_nW_{n+1}, n \in [-3, 2]$
POS features	(1.5) $P_n, n \in [-3, 3]$ (1.6) $P_nP_{n+1}, n \in [-3, 2]$
Dictionary features	(1.7) $D_n, n \in [-2, 2]$ (1.8) $D_nD_{n+1}, n \in [-2, 1]$ (1.9) $D_{-1}D_{+1}$

We incorporate domain knowledge that can be well formulated into first-order logic to extract entity candidates from CRF results. Then, the Markov Logic Networks (MLNs), an undirected graphical model for *statistical relational learning*, is used to validate and correct the errors made in the base model.

MLNs conduct *relational learning* by incorporating first-order logic into probabilistic graphical models under a single coherent framework (Richardson and Domingos, 2006). Traditional first-order logic is a set of hard constraints in which a world violates even one formula has zero probability. The advantage of MLNs is to soften these constraints so that when the fewer formulae a world violates, the more probable it is. MLNs have been applied to tackle the problems of gene interaction discovery from biomedical texts and citation entity resolution from citation texts with state-of-the-art performance (Riedel and Klein (2005), Singla and Domingos (2006)).

We use the Alchemy system (Beta version) (Kok *et al.*, 2005) in our experiment, which is a software package providing a series of algorithms for statistical relational learning and probabilistic logic inference, based on the Markov logic representation.

3.1 Domain Knowledge

We extract 165 location salient words and 843 organization salient words from Wikipedia and the LDC Chinese-English bi-directional NE lists compiled from Xinhua News database. We also make a punctuation list which contains 18 items and some stopwords which Chinese NEs cannot contain. We extract new NE candidates from the CRF results according to the following consideration:

- 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.

- 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.

Table 2: Statistics of NER training and testing corpora.

Corpus	Training NEs	PERs/LOCs/ORGs	Testing NEs	PERs/LOCs/ORGs
CityU	66255	16552/36213/13490	13014	4940/4847/3227
MSRA	37811	9028/18522/10261	7707	1864/3658/2185

NEs: Number of named entities; PERs: Number of person names;
LOCs: Number of location names; ORGs: Number of organization names.

Table 3: OOV Rate of NER testing corpora.

Corpus	Overall (IVs/OOVs/OOV-Rate)	PER (IVs/OOVs/OOV-Rate)	LOC (IVs/OOVs/OOV-Rate)	ORG (IVs/OOVs/OOV-Rate)
CityU	6660/6354/0.4882	1062/3878/0.7850	3947/900/0.1857	1651/1576/0.4884
MSRA	6056/1651/0.2142	1300/564/0.3026	3343/315/0.0861	1413/772/0.3533

IVs: number of IV (named entities in vocabulary); OOVs: number of OOV
(named entities out of vocabulary); OOV-Rate: ratio of named entities out of vocabulary.

- 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.

- For a chunk with low conditional probabilities, all the above assumptions are adopted.

3.2 First-Order Logic Construction

All the above domain knowledge can also be formulated as first-order logic to construct the structure of MLNs. First-order formulae are recursively constructed from atomic formulae using logical connectives and quantifiers. Atomic formulae are constructed using *constants*, *variables*, *functions*, and *predicates*.

For example, we use the predicate organization(candidate) to specify whether a candidate is an ORG. If “中国政府/China Government” is mis-tagged as a LOC by the CRF model, but it contains the organization salient word “政府/Government”. The corresponding formula endwith(r, p)∧orgsalientword(p) ⇒organization(r) means if a tagged entity r ends with an organization salient word p, then it is extracted as a new ORG entity. Typically only a small number (e.g., 10-20) of formulae are needed. We declare 14 *predicates* and 15 first-order formulae according to the domain knowledge mentioned in Section 3.1.

3.3 Training and Inference for Named Entity Correction

Each extracted new 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.

We extract all the NEs from the official training corpora, and then convert them to the first-order logic representation according to the domain knowledge. The MLN training database that consists of predicates, constants, and ground atoms was built automatically. We also extract new entity candidates from CRF results and construct MLN testing database in the same way.

During MLN learning, each formula is converted to Conjunctive Normal Form (CNF), and a weight is learned for each of its clauses. These weights reflect how often the clauses are actually observed in the training data. Inference is performed by grounding the minimal subset of the network required for answering the query predicates. Conducting maximum a posteriori (MAP) inference which finds the most likely values of a set of variables given the values of observed variables can be performed via approximate solution using Markov chain Monte Carlo (MCMC) algorithms. Gibbs sampling can be adopted by sampling each non-evidence variable in turn given its Markov blanket, and counting the fraction of samples that each variable is in each state.

4 Experiment Details

4.1 Data and Preprocessing

The training corpora provided by the SIGHAN bakeoff organizers were in the CoNLL two column format, with one Chinese character per line and hand-annotated named entity chunks in the second column. The CityU corpus was traditional Chinese. We converted this corpus to simplified Chinese and we used UTF-8 encoding in all the experiments so that all the resources (e.g., word dictionary and named entity dictionary) are compatible in our

Table 4: Official results on CityU and MSRA open tracks.

	Precision	*Recall*	$F_{\beta=1}$
CityU			
PER	97.21%	95.26%	96.23
LOC	92.35%	93.42%	92.88
ORG	88.05%	66.44%	75.73
Overall	93.42%	87.43%	90.33
MSRA			
PER	98.33%	94.58%	96.42
LOC	93.97%	93.36%	93.66
ORG	92.80%	84.39%	88.40
Overall	94.71%	91.11%	92.88

system.

Table 2 shows the statistics of NER training and testing corpora and Table 3 shows the OOV (Out of Vocabulary) rate of NER testing corpora [1]. The number of NEs in CityU corpus is almost twice as many as that in MSRA corpus. The OOV rate in CityU corpus is much higher than in MSRA corpus for PERs, LOCs and ORGs. These numbers indicate that NER on CityU corpus is much more difficult to handle.

4.2 Model Development

We performed holdout methodology to develop our model. We randomly selected 5000 sentences from CityU training corpus for development testing and the rest for training. We did the same thing for MSRA training corpus.

To avoid overfitting for CRF model, we penalized the log-likelihood by the commonly used zero-mean Gaussian prior over the parameters. Also, 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).

We found an optimal value for the parameter c [2] for CRFs. Using held-out data, we tested all c values, $c \in [0.2, 2.2]$, with an incremental step of 0.4. Finally, we set $c = 1.8$ for CityU corpus and $c = 1.0$ for MSRA corpus.

5 Official Results

Table 4 presents the results obtained on the official CityU and MSRA test sets. Our results are consistently good: we obtained the first place on the CityU open track (90.33 overall F-measure) and fourth place on the MSRA open track (92.88 overall F-measure) respectively. The lower F-measure obtained on CityU corpus can be attributed to the higher OOV rate of this corpus.

6 Conclusion

We have described a Chinese NER system incorporating probabilistic graphical models and first-order logic which achieves state-of-the-art performance on the open track of SIGHAN-6. We exploited domain knowledge which can capture the essential features of Chinese NER and can be concisely formulated in MLNs, allowing the training and inference algorithms to be directly applied to them. Our proposed framework can also be extendable to NER for other languages, due to the simplicity of the domain knowledge we could access.

References

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.

Stanley Kok, Parag Singla, Matthew Richardson, and Pedro Domingos. The Alchemy system for statistical relational AI. Technical report, Department of Computer Science and Engineering, University of Washington, Seattle, WA, 2005. http://www.cs.washington.edu/ai/alchemy.

Taku Kudo. CRF++: Yet another CRF tool kit. http://crfpp.sourceforge.net/, 2005.

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.

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

Sebastian Riedel and Ewan Klein. Genic interaction extraction with semantic and syntactic chains. In *Proceedings of the Learning Language in Logic Workshop (LLL-05)*, pages 69–74, 2005.

Yanxin Shi and Mengqiu Wang. A dual-layer CRFs based joint decoding method for cascaded segmentation and labeling tasks. In *Proceedings of IJCAI-07*, pages 1707–1712, Hyderabad, India, 2007.

Parag Singla and Pedro Domingos. Entity resolution with Markov logic. In *Proceedings of ICDM-06*, pages 572–582, Hong Kong, 2006.

Xiaofeng Yu, Wai Lam, and Shing-Kit Chan. A framework based on graphical models with logic for Chinese named entity recognition. In *Proceedings of IJCNLP-08*, Hyderabad, India, 2008. To appear.

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.

[1] The NER on the PKU corpus was cancelled by the organizer due to the tagging inconsistency of this corpus.

[2] This parameter trades the balance between overfitting and underfitting. With larger c value, CRF tends to overfit to the give training corpus. The results will significantly be influenced by this parameter