The Sequence Prediction Model of Latent Variable Conditional Random Fields

Jimmy Ming-Tai Wu[1], Jerry Chun-Wei Lin[2,*], Yinan Shao[3], and Matin Pirouz[4]
[1]*College of Computer Science and Engineering*
Shandong University of Science and Technology, Qingdao, China
[2]*Department of Computing, Mathematics, and Physics*
Western Norway University of Applied Sciences, Bergen, Norway
[3]*Department of Computer Science and Technology*
Harbin Institute of Technology (Shenzhen), Shenzhen, China
[4]*Department of Computer Science*
California State University, Fresno, USA
wmt@wmt35.idv.tw; jerrylin@ieee.org; shaoyn0817@163.com; mpirouz@ieee.org

***Abstract*—This paper concentrates on the labeling of sequences and introduces a latent model of CRFs. The latent variable conditional random field uses a latent variable encoding scheme to record the latent structure of hidden variables and observation data. Our proposed model is evaluated on three tasks of sequence prediction, including recognition of the name (NER), chunking, and reference parsing. Experimental results indicate a latent variable in the proposed model and provide competitive and solid performance on all three sequence predictions.**

***Index Terms*—Latent variable CRF, sequence prediction, encoding schema, natural language processing.**

I. Introduction

The critical step in processing text data format is often sequence prediction. The task is to identify each unit/subsequent entry sequences by assigning a semantic label. It can help machines understand the components or structures of certain contexts more effectively. The prediction of a sequence normally comes from a identification of a name entity which extracts names (i.e., person, business, etc.) from texts and from the chunking which identifies the components of the sentences (i.e., names, verbs, adjectives, etc.) and a parsing references which extracts information from a given reference string (i.e., the author, title, journal, etc.). Sequence prediction is a fundamental investigation into the processing of the natural language. Due to its important role in downstream tasks including the extraction relationship [1], [2] and the link-in entity [3], the co-reference resolution [4], it has received considerable attention for several decades. Conventional sequence labeling models, like CRF and the maximum entropy model (MEM), provide the conditional probability model with a representation of the input unit, i.e., characters or words, over the input sequence. Segmentation models, for example, the random Semi-Markov fields (semi-CRF) directly represent the input sequence text span. We will focus on the CRF model in this paper. In [5], [6], it shows that the model performance is affected by the encoding scheme. The BIO and the BILOU are the most commonly used encoding schema where **B** is the beginning, **I** is a denotation inside, **O** is outside, **L** is last and **U** is a unit.

Different coding patterns can lead to various models and sequence-labeling tasks in relation to different performance. In view of a specific model and sequence-labeling task, the validation system is commonly used in order to achieve the best encoding schema. Therefore, in this paper, we propose a latent variable CRF that is able to select the best encoding scheme for each input phrase. The first advantage of this paper is that the developed model uses validation to select encoding schemes manually that our latent variable CRFs can easily and automatically perform. The second advantage of the developed model is that the accuracy performance of all other information (i.e., functional engineering work) can be greatly improved as the model chooses the appropriate coding schemes for each phrase rather than adopting a certain encoding scheme for all phrases. The main contributions of this paper are as follows.

- We introduce a latent variable CRF for the task of sequence prediction, which can be utilized in different sequence prediction subtasks, including name entity recognition, chunking and reference parsing.
- The scheme is used to capture the hidden structures of the hidden variables and the observational data as the latent variables in a proposed model.
- Empirically, we demonstrate a stable impact on performance when you choose the best encoding scheme. The performance of the developed latent variable model is considerably better than the conventional BIO or BILOU encoding schema.

II. Literature Review

The hidden Markov Model (HMM) [7]–[9], the max-entropy (MEM) [10], the conditional random fields (CRF) [11] and the semi-Markov rander-only field (semi-CRF) [12] are all traditional mention-extraction. These are linear models which, given an input sequence, can capture links between labeling within neighborhoods and jointly decode the best label chain. The hidden Markov model (HMM), which can be represented

*Corresponding author

as a dynamic Bayesian network, proposed by Baum and the coworkers [7]–[9]. Fine et al. [13] suggested a Markov-hidden hierarchical model (HHMM), a recursive hierarchical generalization of the Markov-hidden vanilla model. Zhang et al. [14] has built up ICTCLAS, which incorporates in a comprehensive theoretical framework the hierarchical hidden Markov model of Chinese word sequence, part-of-speech tagging, disambiguation, and unknown-word recognition. A general Markov-based model named-entity recognizer was proposed by Shen et al. [15] in the biomedical field. Berger et al. [10] were a pioneer of natural language processing for maximum entropy (MEM).

The maximum Markov entropy (MEMM) model is proposed by McCallum et al. [16] and is a graphical sequence predicting model that combines HMM and MEM features. The problem of continuously using MEM features was investigated by Yu et al. [17]. They explained why MEM worked well with binary functions with the current constraint (MEMC), but not continuous functions. Ratnaparkhi [18] has suggested a statistical model which trains with part-of-the-speech tags and allocates the text with high accuracy that has previously not been visible. They showed that the effectiveness of specialized features to model difficult tagging decisions, and proposed a training strategy that mitigates the corpus-consistency problems discovered during the implementation of specialized features. The Mix-of-Parents Markov Model (MoP-MEMM) is then proposed by Rosenberg et al. [19]. This enables long-range dependencies to be tractable between nodes by limiting the conditionally distributed parent distribution mix of each node.

Lafferty et al. [11], a class of statistical modeling methods, were proposed to be used for random conditions (CRF) models in sequence prediction issues. For these tasks, it offers several benefits over covered Markov modeling and stochastic grammar, including the ability to relax strong assumptions on independence made in such models. The Chinese word segmentation system (CWS) based on model CRF was introduced by Tseng et al. [14]. Zhao et al. [20] considered the CWS problem to be a tagging issue based on character within a conditions based on random fields. They considered a feature template and Tag-Set selection, instead of a method focusing only on a feature template as in the previous work. Based on the selected tag sets, they showed a significant performance difference.

Cuong et al. [21] considered incorporating a high-ordre dependence in conditional random fields between labels or segments. Sarawagi and Cohen [12] proposed semi-markov conditional random fields (Semi-CRF). Importantly, semi-CRF characteristics may measure the segmental characteristics and non-Markovian transitions in a segment. Okanohara et al. [22] introduced techniques that apply semi-CRFs with tractable calculation costs to named tasks of entity recognition. Nguyen et al. [23] extended semi-CRFs of the first-order semi-Markov to include semi-Markov features of the higher order with effective inference and studying algorithms assuming that semi-Markov features of the higher-order are sparse. Muis and Wei [24] have proposed a weak Markov semi-phrase chunking with conditional random fields. For conventional semi-CRF, the model decides intuitively the next segment length and type, while the model tries to suggest a weakened variant in a weak semi-CRF, which takes the two separate decisions by limiting every node to connect either to the label nodes in the next segment, or to all nodes only in the next word. The weak half-CRF model performs much more efficiently, but runs much faster, similarly to conventional semi-CRFs.

In sequence labeling, deep learning methods show benefits. Huang et al. [25] proposed a wide range of long-term sequence prediction models (LSTM) including LSTM, bidirectional LSTM (Bi-LSTM), LSTM with a CRF layer (LSTM-CRF), and bidirectional LSTM with a CRF layer (Bi-LSTM-CRF). Their model produces state of the art POS, chunking and NER data sets with less word embedding compared to earlier observations. Liu et al. [26] proposed a random semi-Markov neural condition field, comprising both the input unit and segment embedding. They conducted experiments in Chinese word segmentation and name entity recognition (NER). Ma and Hovy [27] proposed a CNN-LSTM-CRF model that benefits both from the representation of word and character. They have an end-to-end system, and no pre-processing of features or data is needed. Rei et al. [28] have incorporated character-level information in a sequence prediction to manage the non-vocabulary (OOV) problem. They examined the extensions of the standard LSTM-CRF structure model characteristics. In combination with pre-training word embedding, the encoded character level information was combined with a mechanism that enabled the model to decide dynamically how much information to use from a word or character level component.

Different approaches based on latent variable models previously achieved a sequence prediction. The latent discriminatory model, called Latent Semi-CRF, proposed by Sun and Nan [29], which includes the advantages of two modeling approaches: latent, dynamic CRF and semi-CRF, model the substructure of the sequence of classes, and learn the dynamic among the class labels for the detection of Chinese basic phrases. The discriminative latent variable approach to syntactic parsing reported by Petrov and Dan [30] exists in several levels of refinement. A model of this type, learned by splitting gram marks, is formally a latent variable CRF grammar over trees. A latent semi-CRF model is proposed by Sun et al. [31] to synchronize with their POS the novel words with their POS, regardless of the Chinese text type, without pre-segmentation. Sun and Tsujii [32] described the Latent-Dynamic Inference (LDI), which results in an optimal label sequence with effective search strategies and dynamic programming for the latent conditional models.

III. Preliminaries and Problem Statement

This section presents the preliminary statements and problems of our work, briefly.

A. Latent variable CRF

Consider a sequence of observations $x = (x_1, \dots, x_n)$. The model must determine in the latent variable CRF how to assign a label sequence. $y = (y_1, \dots, y_n)$, from one finite set of labels *Y*. Instead of directly modeling $P(y|x)$, as a conventional CRF would do, a set of latent variables *h* is "inserted" between the *x* and *y* using the chain rule of probability, i.e.,

$$P(y|x) = \frac{1}{Z(x)} \sum_{h} P(y|h, x) P(h|x), \tag{1}$$

where $Z(x)$ is the normalization factor, *h* denotes the latent variable, *x* is the sequence of observations, and *y* represents the sequence of labels. The latent structure between observations and labels can be traced from this model. These models find computer vision applications, specifically video stream recognition and sequence prediction.

B. Encoding Schema

The most popular encoding scheme is BILOU encoding. In the case of the BILOU encoding scheme 1 showing the begin of a segment by **B**, inside a segment by **I**, the last word in a segment by **i** by the end word and the end of word **i** by the word without any segment.

BILOU encoding represents the most popular encoding schema. Fig. 1 shows an example for the BILOU encoding schema, where **B** denotes the beginning of a segment, **I** denotes the inside of a segment, **L** denotes the last word of a segment and **O** stands for the word which does not belong to any segment. 'Michel' is the beginning word of an individual entity as shown in Fig. 1, so marked **B-P**, 'Jordan' is an individual's final word and therefore is marked with **L-P**. The word 'would' is marked **O** because it belongs to no entity. It is a person entity with unit length for the word 'Bush' and therefore it is marked **U-P**. More features are captured by the encoding schema in comparison with the sequence model without any encoding schemes, which has a positive impact on model performance.

C. Problem Statement

Consider an input sequence $x = (x_1, \dots, x_k)$ of length *k*, a label of *x* is a tuple (u, y), which indicates that the *u*-th input word is related to label *y*. A label sequence of *x* is then defined as $s = (s1, \dots, sk)$, where $s_j = (u_j, y_j)$. It is important to note that *x* and *s* input sequences have the same length. In view of the *x* input sequence, the problem of sequence prediction is set of finding *s* of *x* as the most likely label sequence.

IV. Proposed Latent Variable CRF Model

The proposed latent variable CRF model is presented in this section. To give clear reasons, the conventional CRF model is introduced briefly, the proposed latent variable CRF model is specified. The latent variable CRF was designed to utilize the path in the BILOU encoding scheme. This improves prediction results since it can determine the best encoding scheme for each word.

This section introduces the proposed latent variable CRF models. To provide a clear explanation, we briefly introduce the conventional CRF model, then state the proposed latent variable CRF model and explain the main difference between these two models. The designed latent variable CRF (named as LVCRF) is a word level model which hybrids the path in the BIO encoding schema and the path in the BILOU encoding schema, thus enhancing the prediction results than the first approach since the best encoding schema for every word can be determined. Details are respectively described as below. Details are described in the following, respectively.

The CRF is a popular model for the labeling of sequences. CRF can easily incorporate flexible features and manage the problem of label bias in MEM models, in comparison to other models like the hidden Markov or the maximum entropy model (MEM). The latent CRF model has been designed, as it chooses the encoding schema automatically for each word.

The designed latent variable CRF model is a word level model, since it will automatically choose encoding schema for each word. The proposed CRF latent variable is shown by the Fig. 2, which comprises two parts of the proposed model.

The developed latent variable CRF graph model hybridizes the path in the BILOU encoding schema to the specific input sentence. The structure selects the encoding scheme in word rather than sentence level. There are marked paths of the 2^n in the designed model, where n is the length of the sentence. All these paths are identical, such as 'Michel Jordan" and "Bush" are as name entity and label "would choose" as the non-name entity.

The proposed model provides, in the decoding step, a subset of red lines, in which lines are linked together. For example, 'Michel' may be labeled as **B-P** node, while 'Jordan' may be labeled as **L-P** node. For a given input sentence, the designed model determines the best encoding schema for every word.

A. Training Procedure

Following the CRF, we adopted a log-linear approach for such a latent variable CRF. Specifically, for the given input sentence x, the probability of predicting a possible output sequence y reads:

$$p(y|x) = \frac{exp(w^T f(x, y))}{\sum_{y'} exp(w^T f(x, y'))}, \tag{2}$$

where $f(x, y)$ is the feature vector defined over the input-output pair (x, y), and the weight vector w gives the parameters of the model. Our objective is to minimize the regularized negative joint log-likelihood of the dataset, as:

$$L(w) = \sum_i log \sum_{y'} exp(w^T f(x_i, y')) - \sum_i w^T f(x_i, y) + \lambda w^T w, \tag{3}$$

where (x_i, y_i) refers to the *i*-th training instance, and the last term is a $L2$ regularization term with λ setting to 0.01. This objective function may be optimized with standard

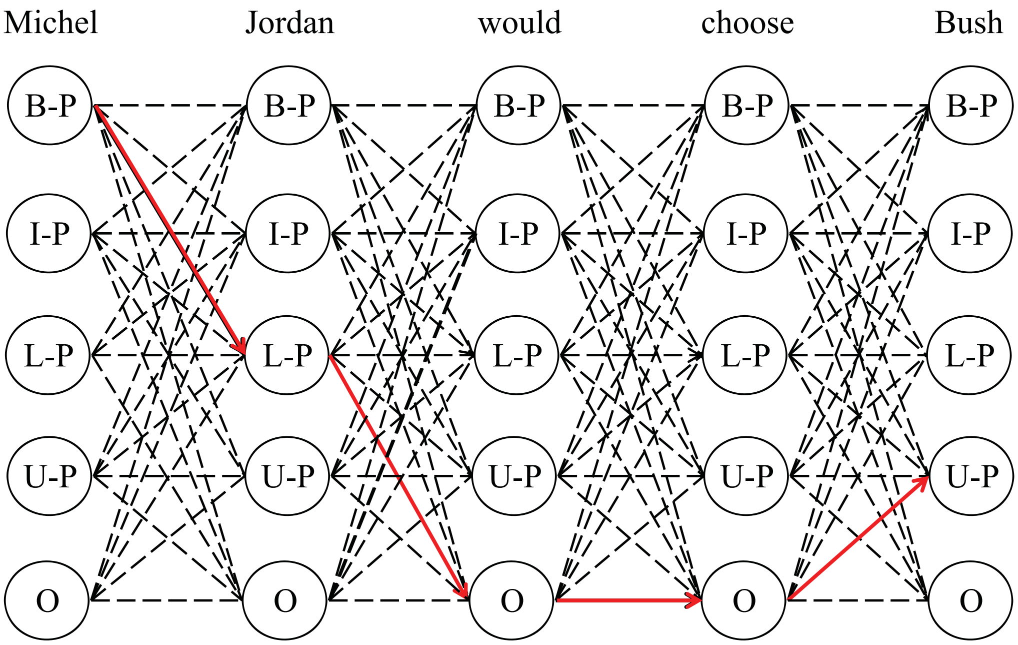

Fig. 1: A encoding BILOU schema.

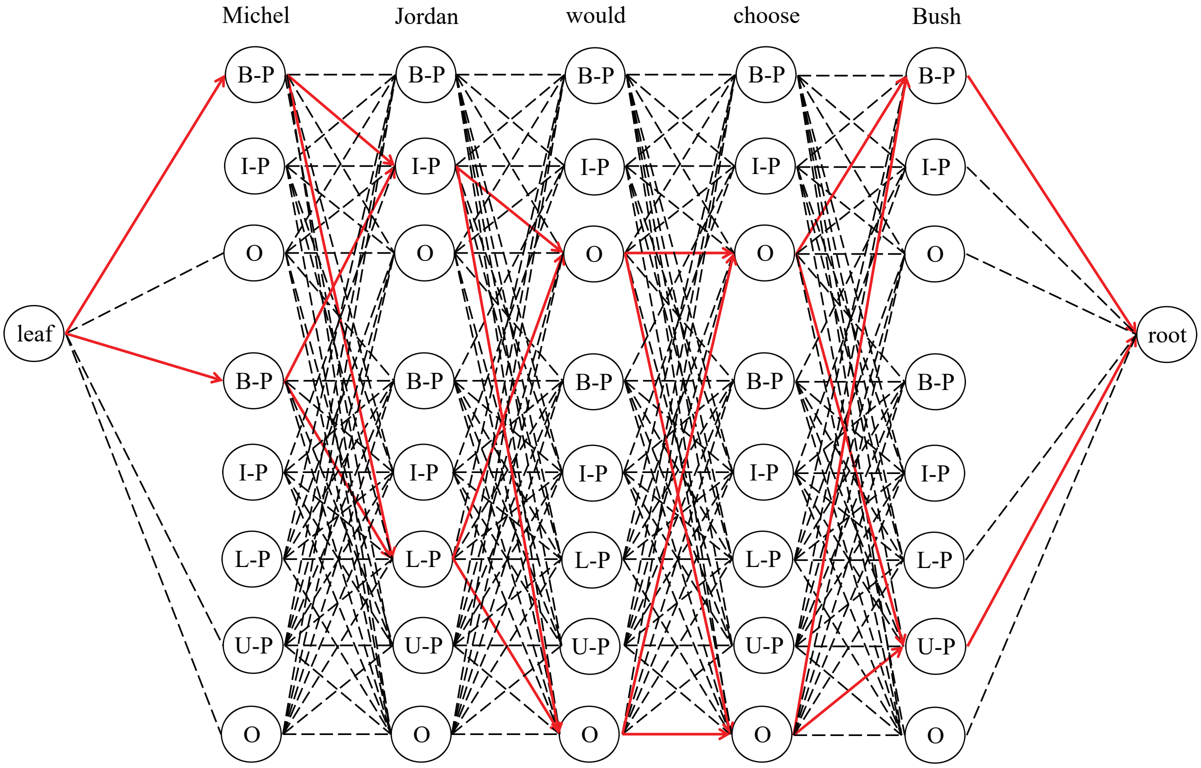

Fig. 2: Proposed latent variable CRF.

gradient-based methods. We use the L-BFGS algorithm as our optimization method and Viterbi algorithm as our inference method in this work.

B. Features

Let us we briefly introduce the CRF feature to compute the **G(x,s)**. Specifically, we consider the following features defined over the inputs.

- **Words features:** Words with a size 3 window appearing around the current position.
- **POS tags features:** POS tags with a size 3 window appearing around the current position.
- **Word *n*-grams features:** Word *n*-grams that contain the position, for n = 2, 3, 4.
- **POS *n*-grams features (if available):** POS tags that contain the current position, for n = 2, 3, 4.

All of these features are utilized for comparison in all conventional CRF models and the proposed latent variable CRF model.

V. Experimental Evaluation

In this section, we evaluate our model on three processing tasks for natural languages: name entity recognition, chunking and reference parsing. The proposed latent variable CRF is thoroughly compared to conventional CRF with BIO and BILOU encoding schemas. The designed model is based on the same feature as the IV-B section. The CRF-BILOU is a BILOU encoding schema with a classical CRF. The models are evaluated as follows.

A. Data Sets

The performance of the designed and the comparison models can be evaluated by using standard data sets in this section. The statistics of the data collection contained in the Table I. The following details are described.

TABLE I: Statistics of the used datasets.

Name	**Task**	**# labels**	**# train**	**# dev**	**# test**
CoNLL2003	NER	8	14,987	3,466	3,684
CoNLL2000	Chunking	22	8,936	N/A	2,012
Cora	Ref parsing	13	500	N/A	N/A

B. Name entity recognition (NER)

Table II compares the performance of different models. The performance of the designed model is better than the existing works. It can be shown that, for a give input sentence, the designed model can effectively choose best encoding schema for every word.

TABLE II: Results on the Conll2003 dataset.

NER task	**Precision**	**Recall**	**F1**
CRF-BIO	84.10	83.59	83.84
CRF-BILOU	83.82	84.36	84.09
Designed model	84.15	85.05	**84.59**

C. Chunking

Table III compares the performance of different models for the chunking task on CoNLL2000 shared task. Results showed that the proposed model outperforms the baseline CF-BIO and CRF-BILOU models. The CRF with BIO encoding schema performs better in the chunking task, while the CRF with BILOU encoding schema outperforms in the name entity recognition. This is due to the fact that none of the encoding schema is the best, it is necessary to choose different encoding schema for different input sentences, as what the designed model does.

TABLE III: Results on the CoNLL2000 dataset.

Chunking task	**Precision**	**Recall**	**F1**
CRF-BIO	90.15	89.89	90.01
CRF-BILOU	90.05	89.88	89.96
Designed model	90.08	90.41	**90.24**

D. Reference Parsing

Table IV compares the performance of the methods for reference parsing on the Cora dataset [33]. For two compared baseline models, we may see that the CRF-BILOU outperforms the CRF-BIO. This may be due to the fact that the CRF-BILOU is capable of capturing more segmental level information, which is quite critical in such tasks. The performance of the proposed model is quite robust, and both outperforms the other two baseline models.

TABLE IV: Results on the CoNLL2000 dataset.

Reference parsing task	**Precision**	**Recall**	**F1**
CRF-BIO	77.92	80.61	79.24
CRF-BILOU	78.35	81.21	79.75
Designed model	78.25	81.89	**80.02**

VI. Conclusion

This paper focuses on the sequence forecasting task. The developed model is utilized to choose the best encoding scheme for every word. Empirically, several standard sequence prediction tasks and datasets have demonstrated the efficiency of the developed model.

References

[1] M. Mintz, R. S. S. Bills, and D. Jurafsky, “Distant supervision for relation extraction without labeled data,” in *The Annual Meeting of the Association for Computational Linguistics*, 2009, pp. 1003–1011.

[2] P. Gupta and B. Andrassy, “Table filling multi-task recurrent neural network for joint entity and relation extraction,” in *The International Conference on Computational Linguistics*, 2016, pp. 2537–2547.

[3] M. W. C. S. Guo and E. Kiciman, “To link or not to link? a study on end-to-end tweet entity linking,” in *The Conference of the North American Chapter of the Association of Computational Linguistics*, 2013, pp. 1020–1030.

[4] J. Lu, D. Venugopal, V. Gogate, and V. Ng, “Joint inference for event coreference resolution,” in *The International Conference on Computational Linguistics*, 2016, pp. 3264–3275.

[5] L. Ratinov and D. Roth, “Design challenges and misconceptions in named entity recognition,” in *The Conference on Computational Natural Language Learning*, 2009, pp. 147–155.

[6] P. L. H. Dai, Y. Chang, and R. T. Tsa, "Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization," *Journal of cheminformatics*, vol. 7(S-1), pp. 1–10, 2015.

[7] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state markov chains," *The Annals of Mathematical Statistics*, vol. 37(6), pp. 1554–1563, 1966.

[8] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology," *Bulletin of the American Mathematical Society*, vol. 37(3), pp. 360–363, 1967.

[9] L. E. Baum, "An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process," *Inequalities*, vol. 3, pp. 1–8, 1972.

[10] A. L. Berger, S. A. D. Pietra, and V. J. D. Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22(1), pp. 39–71, 1996.

[11] J. D. Lafferty, A. Mccallum, and F. C. N. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *The Eighteenth International Conference on Machine Learning*, 2001, pp. 282–289.

[12] S. Sarawagi and W. W. Cohen, "Semi-markov conditional random fields for information extraction," in *The Neural Information Processing Systems*, 2004, pp. 1185–1192.

[13] Y. S. S. Fine and N. Tishby, "The hierarchical hidden markov model: analysis and applications," *Machine Learning*, vol. 32(1), pp. 41–62, 1998.

[14] H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning, "Sequential labeling with latent variables," in *The Workshop on Chinese Language Processing*, 2015, pp. 168–171.

[15] D. Shen, J. Zhang, G. Zhou, J. Su, and C. L. Tan, "Effective adaptation of a hidden markov model-based named entity recognizer for biomedical domain," in *The Nature Language Processing in Biomedicine*, 2003, pp. 49–56.

[16] A. McCallum, D. Freitag, and F. C. N. Pereira, "Maximum entropy markov models for information extraction and segmentation," in *The International Conference on Machine Learning*, 1999, pp. 591–598.

[17] D. Yu, L. Deng, and A. Acero, "Using continuous features in the maximum entropy model, pattern recognition letters," *Pattern Recognition Letters*, vol. 30(14), pp. 1295–1300, 2009.

[18] A. Ratnaparkhi, "A maximum entropy model for part-of-speech tagging," in *The Conference on Empirical Methods in Natural Language Processing*, 1996, pp. 133–142.

[19] D. S. Rosenberg, K. Dan, and B. Taskar, "Mixture-of-parents maximum entropy markov models," in *http://arxiv.org/abs/1206.5261*, 2012.

[20] H. Zhao, C. N. Huang, M. Li, and T. Kudo, "An improved chinese word segmentation system with conditional random field," in *The Workshop on Chinese Language Processing*, 2006, pp. 162–165.

[21] N. V. Cuong, W. S. L. N. Ye, and L. C. Hai, "Conditional random field with high-order dependencies for sequence labeling and segmentation," *Journal of Machine Learning Research*, vol. 15(1), pp. 981–1009, 2014.

[22] D. Okanohara, Y. Miyao, Y. Tsuruoka, and J. Tisuji, "Improving the scalability of semi-markov conditional random fields for named entity recognition," in *The Annual Meeting of the Association for Computational Linguistics*, 2006, pp. 465–472.

[23] V. C. Nguyen, W. S. L. N. Ye, and L. C. Hai, "Semi-markov conditional random field with high-order feature," 2011, pp. 1–4.

[24] A. O. Muis and W. Lu, "Weak semi-markov crfs for noun phrase chunking in informal text," in *The North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 714–719.

[25] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," in *http://arxiv.org/abs/1508.01991s*, 2015.

[26] Y. Liu, W. Che, J. Guo, Q. Bin, and T. Liu, "Exploring segment representations for neural segmentation models," in *The International Joint Conference on Artificial Intelligence*, 2016, pp. 2880–288.

[27] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," in *The Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 1064–1074.

[28] M. Rei, G. K. O. Crichton, and S. Pyysalo, "Attending to characters in neural sequence labeling models," in *http://arXiv:1611.04361*, 2016.

[29] X. Sun and X. Nan, "Chinese base phrases chunking based on latent semi-crf mode," in *The International Conference on Natural Language Processing and Knowledge Engineering*, 2010, pp. 1–7.

[30] S. Petrov and K. Dan, "Sparse multi-scale grammars for discriminative latent variable parsing," in *The Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 867–876.

[31] X. Sun, D. Huang, and F. Ren, "Detecting new words from chinese text using latent semi-crf models," *IEICE Transactions on Information and Systems*, vol. 93(6), pp. 1386–1393, 2010.

[32] X. Sun and J. Tsujii, "Sequential labeling with latent variables," in *The European Chapter of the Association for Computational Linguistics*, 2009, pp. 772–780.

[33] L. E. Baum and G. R. Sell, "Growth transformations for functions on manifolds," *Pacific Journal of Mathematics*, vol. 27(2), pp. 211–227, 1968.