Labeling Sequential Data Based on Word Representations and Conditional Random Fields

Xiuying Wang, Bo Xu, Changliang Li, and Wendong Ge

***Abstract*—Most of Natural Language Processing tasks including part-of-speech tagging, chunking, named entity recognition can be seen as tasks assigning labels to words. Many existing methods including hidden Markov models, maximum entropy Markov models and conditional random fields have been applied to label sequential data, which rely on amount of training data and can't solve the problem of out-of-lexicon words. In this paper, we propose a new method based on word representations and conditional random fields to solve these problems. We preprocess input features via computing word similarity based on word representations which can capture semantic similarity of words on the basis of vast amounts of unlabeled training data, and then use these preprocessed features as input features of training data to train conditional random fields model. The experiment results show that our approach has improvements in labeling accuracies upon the existing methods.**

***Index Terms*—Conditional random fields, label sequential data, word representations, word similarity.**

I. Introduction

Along with the rapid development of computer technologies and Internet technologies, a great number of data have been being accumulated. Unlike data in database, the vast majority of these data are unstructured, making information extracting a challenging task. In recent years, there is rising interest in transforming the data from unstructured form into structured representation, which can be seen as tasks assigning labels to words. Therefore, it is meaningful task to label sequences.

Many Natural Language Processing (NLP) tasks including part-of-speech tagging (POST), chunking (CHK), named entity recognition (NER) can be seen as tasks assigning labels to words [1]-[5]. Many approaches have proposed for labeling sequential data in the past, including: hidden Markov models (HMMs), maximum entropy Markov models (MEMMs) and Conditional Random Fields (CRFs).

HMMs and stochastic grammars are generative models that assigning a joint probability of the observation data and labelling sequences. Their parameters are typically trained by maximizing the joint likelihood of training data. To define a joint probability of the observation data and label sequences, the generative model needs to enumerate all possible observation sequences, typically requiring a representation in which observations are task-appropriate atomic entities, such as words or nucleotides [6]. Therefore, the disadvantages of HMMs are the need for an a priori notion of the model topology and, as with any statistical technique, large amounts of training data. What's more, HMMs can't take full account of contextual feature due to observations of HMMs in different frames (e.g., word tokens at different positions) are assumed to be independent given the state. MEMMs are conditional probabilistic sequence models that solve problems of HMMs [7]. MEMMs select features at random. In MEMMs, each word token has a exponential model that takes the observation features as input, and outputs a distribution over possible next states. An appropriate iterative scaling method is used to train these exponential models in the maximum entropy framework. Although, MEMMs can solve problems of HMMs and select features at random, it only does normalization in local, making itself fall into local optimum easily and arise the label bias problem. CRFs have been successfully applied in many NLP tasks for several years. CRFs are generative models that estimate the label sequences CRFs probability directly. Unlike HMMs, CRFs do not require the assumption that observations are assumed to be independent, hence CRFs have high flexibility in choosing features. Furthermore, CRFs do normalization in global, solving the label bias problem. However, CRFs have disadvantages of large amounts of training data which are annotated manually, long training time. To solve these problems, many scholars proposed many semi-supervised method to improve CRFs [8], [9]. Dong Yu *et al.* proposed deep-structured CRFs [10]. Meanwhile, there are many approaches which have been proposed for semi-supervised learning in the past, including: generative models [11]-[14], self-learning [15], [16], co-training [17], information theoretic regularization [18], and graph based on transductive methods [18]-[20]. In the task, we can construct field-dependent lexicons that enumerate possible values for each label. However, approach based a pure lexicon-lookup is insufficient largely, since the presence of ambiguous words. Another challenge is that users may formulate sentences using out-of-lexicon words. What's more, the training of most statistical approach relies on amount of training data largely.

For these reasons, we take a statistical approach which incorporated word representations into conditional random fields to label sequential data so that to solve our problems. More specifically, firstly, we use the classifier which is based on word representations [21], [22] to classify all the words of hotel reservation. The aim of this step is to improve labeling accuracies of out-of-lexicon words and reduce the workload of hand-labeled data through preprocessing training data and testing data of hotel reservation. What's more, we use CRFs [6] that model probabilistic dependencies between two consecutive labels and between labels and observations. Although all words have been classified, we can't assign

Manuscript received October 6, 2014; revised August 7, 2015.

The authors are with the Institute of Automation, Chinese Academy of Sciences, China (e-mail: xiuying.wang@ia.ac.cn).

doi: 10.18178/ijmlc.2015.5.6.548

labels to words directly for ignoring context information. Therefore, we use CRFs to label data. Unlike the general methods such as HMMs, MEMMs and CRFs, we employ a classifier based on word representations to preprocess training data and testing data, improving labeling accuracies effectively.

In this paper, we extract the key information of a sentence, making the machine to better understand the meaning of the sentence. Therefore, when sentence is against structured data, it is beneficial to extract information from sentences that is explicitly represented in a structured form. In this paper, we study the problem of sentence tagging as one step toward this goal. More specifically, we view a sentence as a sequence of word tokens. Given a set of predefined labels, our aim is to assign each word token a label indicating what is the key information of a sentence. In particular, we focus our attention on assigning labels to the sentences of hotel reservation, since this is one typical domain where structured information can help make machine to understand what the user said.

Specifically, we make the following contributions:

1) We learn vector representations of words by improving Skip-gram model and Continuous Bag-of-Words (CBOW) model recently proposed by Mikolov *et al*. [23].
2) We improve labeling accuracies of out-of-lexicon words via the method CRFs based on vector representations and CRFs.
3) We can reduce the workload of hand-labeled data through using a classifier based on word representations to preprocess training data.

The remainder of this paper is organized as follows. In Section II, we simply introduce CRFs. Section III describes our proposed method based on vector representations and CRFs for labeling sequential data. Section IV presents the experimental results. In Section V, we conclude ideas and future research.

II. Conditional Random Fields

The linear-chain CRFs showed in Fig. 1 is the most popular CRF for sequential labeling because of its simplicity and efficiency. We choose to apply linear-chain CRFs to our task due to its ability of incorporating arbitrary features functions on observations without complicated the training. Formally, we let $\boldsymbol{x} = (x_1, x_2, \ldots, x_T)$ denote an input query of T-frame observation sequence, and $\boldsymbol{y} = (y_1, y_2, \ldots, y_T)$ denote the corresponding label sequence. Each y_i can take as a pre-defined categorical value. We further augment a state sequence with two special states: Start and End, denoted by y_0 and y_{T+1} respectively.

In the linear-chain CRFs, the conditional probability $p(y \mid x)$ of a label sequence given the observation sequence is given by

$$p(\boldsymbol{y} \mid \boldsymbol{x}; \Lambda) = \frac{1}{Z(\boldsymbol{x}; \Lambda)} \exp\left\{ \sum_k \lambda_k \sum_{t=1}^{t=T+1} f_k(y_{t-1}, y_t, \boldsymbol{x}, t) \right\} \quad (1)$$

where, $f_k(y_{t-1}, y_t, \boldsymbol{x}, t)$ which represent features of the whole observation sequence and the relevant label sequence at time t and time $t-1$ is a transition function. $f_k(y_t, \boldsymbol{x}, t)$ which represent features of the whole observation sequence and the relevant label sequence at time t is a state function. $\Lambda = \{\lambda_i\}$ are the model parameters, and

$$Z(\boldsymbol{x}; \Lambda) = \sum_{\boldsymbol{y}} \exp\left\{ \sum_k \lambda_k \sum_{t=1}^{t=T+1} f_k(y_{t-1}, y_t, \boldsymbol{x}, t) \right\} . \quad (2)$$

The partition function $Z(x; \Lambda)$ normalizes the exponential form so that it becomes a valid probability distribution.

Given a set of manually-labeled samples $\{(x^{(i)}, y^{(i)})\}_i^m$, we can estimate model parameters $\Lambda = \{\lambda_i\}$ in a supervised fashion. In supervised training, the aim of estimating model parameters is to maximize the conditional log-likelihood of training data while regularizing model parameters:

$$J_1 = \sum_k \log p(\boldsymbol{y}^{(k)} \mid \boldsymbol{x}^{(k)}; \Lambda) - \frac{\| \Lambda \|}{2\sigma^2} \quad (3)$$

where, σ^2 is a parameter that balances the log-likelihood and the regularization term. $\| \Lambda \| / 2\sigma^2$ is a standard regularization used to limit over-fitting on rare features and avoid degeneracy in the case of related features.

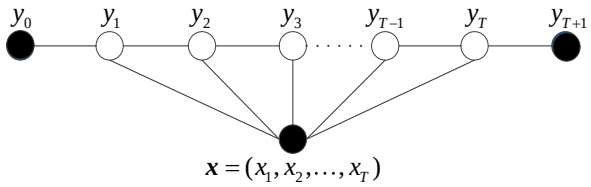


Fig. 1. Graphical model representation of the linear-chain CRFs, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_T)$ is the observation sequence and is the label sequence. The solid nodes denote observed variables, and the empty nodes denote unobserved variables.

III. Vector Representations and CRFs for Labeling Data

In this work, we employ a classifier based on word representations to preprocess training data and testing data which makes generalization to all words of the training data and testing data, and then we trained CRF model label by using preprocessed training data as input features of CRFs. Because the model used to learn vector representations of words is trained on a large corpus, a classifier can be learned through computing similarity threshold between words. Our purpose of doing follow-up steps is to improve labeling accuracies of out-of-lexicon words effectively and reduce the workload of hand-labeled data. For example, *double room* is a word of out-of-lexicon while *single room* is a word of training data, then *double room* can be assigned to the same category in which *single room* is. In this way, we not only reduce the workload of hand-labeled data via using classifier to classify all the words of training data, but also improve labeling accuracies of out-of-lexicon words via classifier for its model is trained on the basis of vast amounts of unlabeled training data. In this paper, we learn vector representations of words by improving Skip-gram model and Continuous

Bag-of-Words (CBOW) model recently proposed by Mikolov *et al*.

A. Word Representation

Distributed representations of words were proposed by Rumelhart *et al*. [21] and have been applied in NLP tasks such as word representation learning, named entity recognition, disambiguation, parsing, and tagging successfully. The vector representations of words promote learning algorithms to achieve better performance in NLP tasks by grouping similar words. Now, there are many typical models for word representation such as (SENNA) [24], hierarchical log-bilinear (HLBL) [25] and recurrent neural network based language model (RNNLM) [26]. In this work, we employ the distributed Skip-gram model and CBOW model to learn word representations because the models can be trained on a large corpus in hours for its simplicity. We will describe the model in detail as follows.

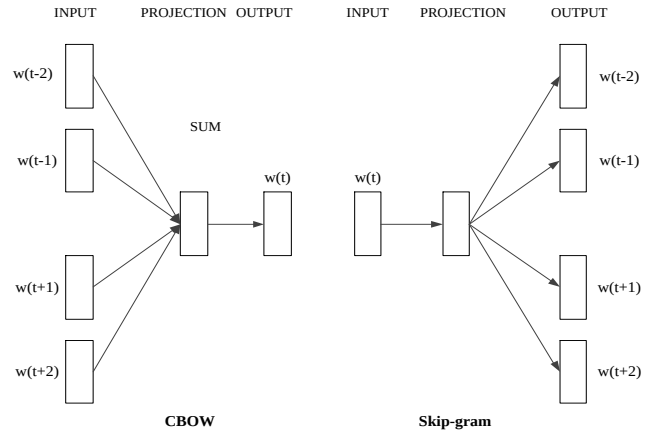


Fig. 2. Graphical representation of the Skip-gram model and CBOW model. The CBOW model predicts the word in the middle, and the Skip-gram model predicts surrounding words in same sentence.

The model architectures are shown in Fig. 2. In practice, the training objective of the Skip-gram is to learn word representations that are good at predicting the surrounding words in same sentence [21].

More formally, given a series of training words $(w_1, w_2, w_3, \ldots, w_T)$, the objective of the Skip-gram model is to maximize the average log probability by using (4).

$$\frac{1}{T}\sum_{t=1}^{T}\left[\sum_{-k\le j\le k, j\ne 0}\log p(w_{t+j}\mid w_t)\right] \quad . \tag{4}$$

where k is the size of the training window. The inner summation ranges $-k$ from k to compute the log probability of correctly predicting the word w_{t+j} given the word w_t in the middle. The outer summation goes through all words of the training corpus. The basic Skip-gram formulation $p(w_{t+j}\mid w_t)$ is defined as

$$p(w_i \mid w_j) = \frac{\exp({u_{w_i}}^{\mathrm{T}} v_{w_j})}{\sum_{l=1}^{V}\exp({u_l}^{\mathrm{T}} v_{w_j})} \tag{5}$$

where u_w and v_w are the "input" and "output" vector representations of w respectively. V is the number of words in the vocabulary.

In this work, in order to better capture semantic similarity of words, we expand center word through semantic dictionary. The model architectures are shown in Fig. 3. Because the surrounding words of similar words are very like be same and synonyms can replace each other in the sentence, we can improve accuracy of word representations in the vector space through adding semantic dictionary to expand center word.

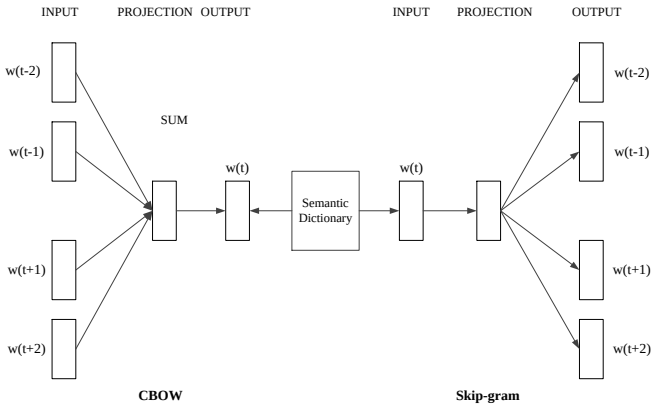


Fig. 3. Graphical representation of the Skip-gram model and CBOW model based on semantic dictionary.

As same as above, given a series of training words $(w_1, w_2, w_3, \ldots, w_T)$ and a series of synonym of the word w_t in the middle $(w_1', w_2', \ldots, w_N')$ (including the word w_t), the objective of improved the Skip-gram model is to maximize the average log probability by using (6).

$$\frac{1}{N}\left(\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\sum_{-k\le j\le k, j\ne 0}\log p(w_{t+j}\mid w_{t+i}')\right]\right) \quad . \tag{6}$$

In this work, because our experimental data are against the sentences of hotel reservation which are all Chinese, we use HIT IR-Lab Tongyici Cilin as semantic dictionary.

B. Architecture of Labeling Method Based on Word Representation and CRFs

Combined with word representation and CRFs, we can obtain a new model for labeling Sequential Data. The model architectures are shown in Fig. 4.

More formally, we preprocess training data through generalizing all words of training data firstly. We represent all words as word vectors with pre-trained 100-dimensional word vectors from unsupervised model. Similar to other local co-occurrence based vector space models, the resulting word vectors capture syntactic and semantic information. They use unlabeled data to induce word representations by predicting how likely it is for a word to occur in the text. In this paper, we use Baidu Encyclopedia text, about 110 million words. What's more, by via computing cosine value of vectors of two words, we can get similarity value between two words. If the training data has words, we can get a symmetric matrix. We can generalize all words of training data by the method of

classifier based on this matrix that we have obtained. Finally, we use generalizes words as the observation sequence of CRF models which we need to train. Combined with pre-trained label sequence, we can learn our model.

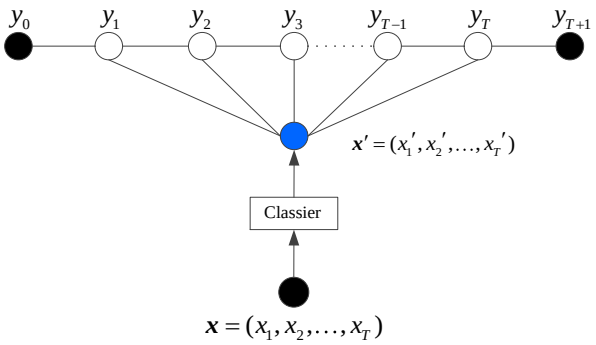


Fig. 4. Graphical model representation based on word representations and CRFs, where $x' = (x_1', x_2', \ldots, x_T')$ is the observation sequence via a classier based on word representations which maps raw observation sequence $x = (x_1, x_2, \ldots, x_T)$ to $x' = (x_1', x_2', \ldots, x_T')$.

Specifically, there are the following advantages of our model:

- The supervised CRF model needs a lot of man-labeled corpus as training data to train model. We can reduce the amount of hand-labeled data through our model based on word representations and CRFs.
- Our model can improve labeling accuracies of out-of-lexicon words. When a word which appears in user input is out-of-lexicon words, we can generalize the word via judging this new word which class it belongs to. And then, our model can label the new words more correctly than user input CRF model.

IV. Experiments

We use two evaluation metrics which are sentence accuracy and word accuracy to evaluate our method. Sentence accuracy is the percentage of sentences that are correctly labeled. Word accuracy is the percentage of word tokens that are correctly labeled. We can use the precision, recall and F-measure given as follows to compare the performance of different taggers learned by different methods.

$$P = \frac{TP}{T} \tag{7}$$

$$R = \frac{TP}{N} \tag{8}$$

$$F = \frac{(\beta^2 + 1)P \times R}{\beta^2 (P \times R)} \tag{9}$$

where P is precision which denotes ratio between correct predicted sentences and predicted sentences. R is recall rate which denotes ratio between correct predicted sentences and pre-defined predicted sentences. F is comprehensive evaluation index of precision and recall rate. In this paper, we set the parameters $\beta = 1$.

A. Task of Hotel Reservation Query Tagging

In the label query task, each query is sequence of word tokens. Our goal is to assign a label from a set of predefined fields to each word token. Fig. 5 is an example showing a sentence of hotel reservation annotated with labels. More specifically, we focus on tagging hotel reservation queries with ten fields given in Table I.

TABLE I: Labels Used in Tagging Task of Hotel Reservation

Labels	Abbreviated	Example
room type	RT	A *single room* available for October 4
room count	RC	*A* single room available for October 4
order time	OT	A single room available for *October 4*
how long	HL	I want to live for *three days*.
client name	NR	My name is *Tom*, telephone number is 88802234.
client phone	CP	My name is Tom, telephone number is *88802234*.
leave time	LT	I leave on *October 4*.
client count	CC	We have *three persons*.
card number	CN	My passcard number is *333331111*.
other	O	A single room *available for* October 4

O O RC RT RT O O OT OT

I want a single room available For October 4

Fig. 5. A sentence of hotel reservation annotated with labels. Italics are words of sentence, the other words are labels.

B. Data

Our task aims at labeling Chinese domain-specific data, however, there is no such standard dataset. Therefore, we collect our experiment data from web, actual dialogue scenes and reservation call which are all about hotel reservation. We can collect 170 dialogues about hotel reservation which contain 4400 sentences. In this work, we only focus on sentences talked by users because they contain useful information we need. Of the remaining sentences, we use 861 randomly sampled ones for training and 200 ones for testing.

C. Results

In our evaluation, we use two different feature sets for all experiments: (1) unigram features only, and (2) unigram + regular expression features. The former feature set completely relies on training data. The latter feature set has a better generalized ability, but it needs engineering efforts from human to design regex rule.

Under the above features, we compare the following two methods for labeling sequential data:

1) Supervised CRF. Train a linear-chain CRF on hand-labeled samples.
2) The model based on word representations and CRFs. The method is introduced in detail in Section III.

Table II lists results of word accuracy for two methods together with different features used by each method, and Table III lists results of sentence accuracy. Row 1 is supervised CRF model with unigram features. Row 2 is supervised CRF model with unigram + regular expression features. Row 3 and row 4 are our proposed method. Row 3 is the model based on word representations and CRFs with unigram features, while row 4 is model based on word representations and CRFs with unigram + regular expression features. There are some clear trends in the results of Table II

and Table III:

1) The model based on word representations and CRFs significantly outperforms CRFs.
2) Incorporating regular expression rule into the models, can significantly improve the labeling accuracies of Sequential Data.

TABLE II: COMPARISON WITH DIFFERENT METHODS FOR LABELING ACCURACIES OF WORDS

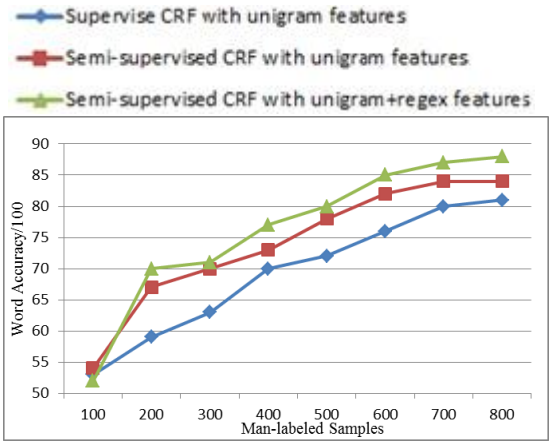
Method	Feature Sets	P	R	F
CRF	unigram features	0.795	0.8368	0.8154
CRF	unigram + regular expression features	0.845	0.8895	0.8667
word representations + CRFs	unigram features	0.825	0.8684	0.8462
word representations + CRFs	unigram + regular expression features	0.87	0.9158	0.8923

TABLE III: COMPARISON WITH DIFFERENT METHODS FOR LABELING ACCURACIES OF SENTENCE

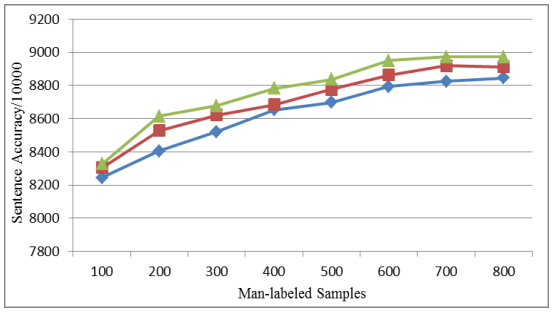
Method	Feature Sets	P	R	F
CRF	unigram features	0.8758	0.89	0.8828
CRF	unigram + regular expression features	0.8906	0.9	0.895
word representations + CRFs	unigram features	0.8877	0.9019	0.8462
word representations + CRFs	unigram + regular expression features	0.8942	0.9107	0.9023

D. Impact of Amount of Training Data

As shown in Fig. 6, we can see: (1) Adding regular expression features better than using unigram features only. (2) The model based on word representations and CRFs performs significantly better than CRF in both feature sets. This confirms that transition features are helpful in labeling sequential data. (3) The model based on word representations and CRFs performs better when amount of training data is larger. (4) Incorporated word vector into CRFs, can reduce the amount of hand-labeled data.



(a) Labeling accuracies of words with different amounts of man-labeled samples



(b) Labeling accuracies of Sentences with different amounts of man-labeled samples

Fig. 1. Labeling accuracies with different amounts of man-labeled samples. Horizontal axis presents the amount of training data, and vertical axis presents accuracy of sentences or word.

V. CONCLUSIONS AND FURTHER WORK

We presented a new model based on word representations and CRFs. Our aim is to improve labeling accuracies effectively and reduce the amount of man-labeled data. Furthermore, we incorporate regular expression features into our models to improve labeling accuracies effectively. Results show that semi-supervised CRFs model can improve labeling performance compared with supervised CRF effectively. In the future, we would like to learn a better word representation to preprocess training data, since word representation is better, the classifier of words is more accurate. It is also important to label sequential data.

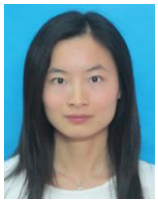
ACKNOWLEDGMENT

The authors would like to thank the three anonymous referees for their insightful comments to the improvement in technical contents and paper presentation.

REFERENCES

[1] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. Int. Conf. Mach. Learn*, 2001, pp. 282-289.
[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. ICLR*, 2013, pp. 1301-3781.
[3] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," in *Proc. ICLR*, 2013, pp. 1309-4168.
[4] T. Grenager, D. Klein, and C. Manning, "Unsupervised learning of field segmentation models for information extraction," in *Proc. 43rd Annu. Meeting ACL*, 2005, pp. 371-378.
[5] X. Li, Y. Y. Wang, and A. Acero, "Extracting structured information from user queries with semi-supervised conditional random fields," in *Proc. SIGIR'09*, July 2009.
[6] X. Li, "On the use of virtual evidence in conditional random fields," in *Proc. EMNLP*, Aug. 2009.
[7] D. Pinto, A. McCallum, X. Wei, and W. B. Croft, "Table extraction using conditional random fields," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2003.
[8] D. Yu and L. Deng, "solving nonlinear estimation problems using splines," *IEEE Signal Process. Mag.*, vol. 26, no. 4, pp. 86-90, 2009.
[9] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy Markov models for information extraction and segmentation," in *Proc. ICML 2000*, 2000, pp. 591-598.

[10] V. Castelli and T. Cover, "The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter," *IEEE Trans. on Information Theory*, vol. 42, no. 6, pp. 2102-2117, 1996.
[11] I. Cohen and F. Cozman, *Risks of Semi-supervised Learning*, MIT Press, 2006, pp. 55-70.
[12] A. McCallum Nigam, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, vol. 39, no. 2, pp. 135-167, 2000.
[13] G. Mann and A. McCallum, "Generalized expectation criteria for semi-supervised learning of conditional random fields," *Proceedings of the Association for Computational Linguistics*, ACL Press, 2008.
[14] G. Celeux and G. Govaert, "A classification EM algorithm for clustering and two stochastic versions," *Computational Statistics and Data Analysis*, vol. 14, pp. 315-332, 1992.
[15] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proc. the 33rd Annual Meeting of the Association for Computational Linguistics*, 1995, pp. 189-196.
[16] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. the Workshop on Computational Learning Theory*, 1998, pp. 92-100.
[17] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Advances in Neural Information Processing Systems*, vol. 17, pp. 529-536, 2004.
[18] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," *Advances in Neural Information Processing Systems*, vol. 16, pp. 321–328, 2004.
[19] D. Zhou, J. Huang, and B. Scholkopf, "Learning from labeled and unlabeled data on a directed graph," in *Proc. the 22nd International Conference on Machine Learning*, 2005, pp. 1041-1048.
[20] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semisupervised learning using Gaussian fields and harmonic functions," in *Proc. the 20th International Conference on Machine Learning*, 2003, pp. 912–919.
[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. ICLR*, 2013, pp. 1301-3781.
[22] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting Similarities among Languages for Machine Translation," in *Proc. ICLR*, 2013, pp. 1309-4168.
[23] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 33-536, 1986.
[24] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493-2537, 2011.
[25] A. Mnih and G. Hinton, "A scalable hierarchical distributed language model," in *Proc. the Conference on Neural Information Processing Systems*, 2008, pp. 1081-1088.
[26] M. Tomą, "Statistical language models based on neural networks," PhD thesis, Brno University of Technology, 2012.



Xiuying Wang was born on January 15, 1986 in Jiangxi. In 2011, she received the master's degree in power electronic and power drivers of North China University of Technology which is located in Beijing, China. Now she is studying for a doctor's degree in pattern recognition at the Institute of Automation, Chinese Academy of Sciences. Her research interests include pattern recognition, neural networks, machine learning, natural language processing, and spoken dialog systems.

Bo Xu is a professor in Institute of Automation, Chinese Academy Sciences (IACAS). He received the B.S. degree in electrical engineering from Zhejiang University in 1988, the M.S. degree in 1992 and Ph.D. degree in 1997 in Institute of Automation, Chinese Academy Sciences. His current research interests include multilingual speech recognition and machine translation, multimedia Internet content intelligent processing, interactive immersive 3D Internet.

Changliang Li was born in 1985. In 2011, he received the B.S. degree in communication science from Northeast University. Now, he is a Ph.D. student in Institute of Automation, Chinese Academy of Sciences which is located in Beijing. His research interest includes natural language processing, big data, deep learning.

Wendong Ge is an assistant professor in Institute of Automation, Chinese Academy Sciences (IACAS). He received the B.S. degree in biomedical engineering from Tianjin Medical University in 2006, the M.S. degree in information and communication engineering from Southeast University (NIT) in 2009, and Ph.D. degree in information and communication engineering from Beijing University of Posts and Telecommunications (BUPT) in 2012. He was involved in several projects funded by the National High Technology Research and Development Program of China, and National Natural Science Foundation of China. From November 2010 to November 2011, he was a joint Ph.D. student at the University of British Columbia, Vancouver, BC, supported by China Scholarship Council. His current research interests include spoken dialogue systems, dialogue management, reinforcement learning and deep learning.