ICSP2008 Proceedings

Chinese Spoken Language Understanding with Conditional Random Fields

Weidong ZHOU *†, Baozong YUAN *, Zhenjiang MIAO *, Weibin ZHU*, Weibin LIU*
** Institute of Information Science, Beijing Jiaotong University, Beijing 100044, P.R.China*
†College of Information and Control Engineering, Petroleum University, Dongying 257061, P.R.China
Email: mlzhou@hdpu.edu.cn, bzyuan@bjtu.edu.cn, zjmiao@bjtu.edu.cn, wbzhu@bjtu.edu.cn, wbliu@bjtu.edu.cn

Abstract

This paper presents a new approach to spoken language understanding tagging for Chinese texts using conditional random fields (CRFs). SLU for spoken dialog system, especially the slot-filling problem is solved as a sequential labeling problem. The experimental results show that this approach has good performance and is feasible for the restricted domain oriented Chinese spoken language understanding.

1. Introduction

Spoken language understanding (SLU) is a key component in spoken dialog system. SLU is aimed at the interpretation of signs conveyed by a speech signal. A user's utterance is firstly converted into a text string which is the input of SLU component by automatic speech recognition (ASR) component. Due to the limitation of ASR technology, the text string may include some errors. On the other hand, spoken language is much noisier than written language [1]. The inputs to an SLU system are not as well formed as those to an NLU system. They often do not comply with rigid syntactic constraints. Disfluencies such as false starts, repairs, and hesitations are pervasive in conversational speech. Traditionally, SLU task has been accomplished by writing context-free grammars (CFGs) or unification grammars (UGs) by hand. The manual grammar authoring process is laborious and expensive, requiring much expertise. In recent years, many data-driven models have been proposed for this problem [2][3][4][5]. Statistical spoken language understanding (SSLU) is a research interest lying in the intersection of SLU and machine learning. SSLU differs from traditional spoken language understanding in that instead of having a linguist manually construct some model of a given linguistic phenomenon, that model is instead (semi-) automatically constructed from linguistically annotated resources. Statistical SLU can automatically learn from training examples with corresponding semantics. While rule-based approach often achieves good performance in commercial systems, data-driven approach is more robust and feasible because it is more portable and less expensive to build semantic labeled data for a new domain. Moreover, the statistical framework shown in Fig.1 is growing into a more powerful tool for SLU by using modern machine learning methods: the system can reduce the labeling efforts by active/semi-supervised learning and can incorporate prior knowledge into statistical model.

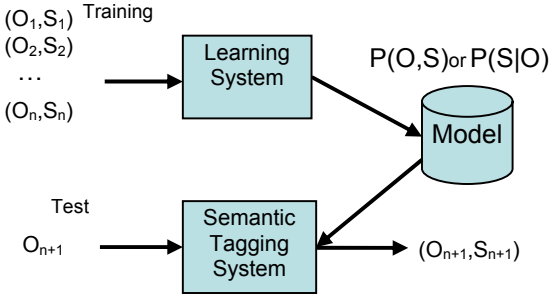


Fig.1 the framework of statistical SLU

In this paper, we think that SLU for spoken dialog system, especially the slot-filling problem may be solved as a sequential labeling problem. A method of using conditional random fields (discriminative model) for semantic tagging is introduced in this paper. This paper is organized as follows. In section 2 we explain why statistical SLU is regarded as a sequence tagging problem. CRF model for sequence labeling is reviewed in Section 3. Preliminary experimental results are discussed in section 4. Concluding comments and future work can be found in section 5.

978-1-4244-2179-4/08/$25.00 ©2008 IEEE

2. Statistical SLU as a sequence tagging problem

For spoken dialog system, SLU aims to fill the domain-specific frame slots from speech recognized utterance. A semantic frame is a well-formed and machine-readable structure of extracted information consisting of slot/value pairs. An example of such a reference frame for train information data is as follows.

<s>五月八号从北京到上海的火车有哪些</s>

(English meaning: <s> which of trains are traveled from Beijing to Shanghai </s>)

Slot name	value
Request Action	火车有哪些 (Request train number)
FROMLOC.CITY_NAME	北京(Beijing)
TOLOC.CITY_NAME	上海(Shanghai)
MONTH_NAME	五月(May)
DAY_NUMBER	八号(eighth)

Tab.1 the slot/value pairs

O:	五月八号	从北京	到上海	的	火车有哪些
S:	Date	Fromloc	toloc		Request action

Tab.2 the observe sequence and semantic sequence

In the statistical framework, the SLU problem can be stated as a sequential supervised learning problem. Let $G = \{O^{(i)}, S^{(i)}\}_{i=1,\cdots N}$ be a set of N training examples. Each example is a pair of sequence $\{O^{(i)}, S^{(i)}\}$, where feature vector sequence $O^{(i)} = (O_1^{(i)}, O_2^{(i)}, \cdots, O_{Ti}^{(i)})$ and label sequence $S^{(i)} = (S_1^{(i)}, S_2^{(i)}, \cdots, S_{Ti}^{(i)})$ co-exist. The goal of statistical SLU is to construct a classifier f that can correctly predict a new label sequence given an input sequence O, i.e., the goal of classifier f is to find the best probable semantic class sequence $\hat{S} = \arg\max_s f(o, s, \Lambda)$.

Note that input sequence o and output sequence s is significantly correlated by a complex structure. In the sequential supervised learning task, the structure is simplified to left-to-right sequential structure. The family of techniques for solving such structured problems is generally known as structured prediction or structured learning in machine learning community. To date, there are several state-of-the-art structured learning algorithms such as hidden Markov model, conditional random fields [6], maximum-margin Markov network [7], support vector machine for structured outputs [8] and search-based structured prediction [9]. The developer has a great freedom to design the function $f(o, s, \Lambda)$.

The traditional way to build sequential model $f(o, s, \Lambda)$ is to use an HMM which represents the joint probability P(O, S). However, modeling the joint distribution is difficult for exploiting the rich local features, because it requires modeling the complex dependencies to estimate distribution P(O). The alternative is a discriminative approach that directly models the conditional distribution P(S|O). Using the discriminative classifier provides great flexibility to include a wide variety of arbitrary, non-independent features of the input. For our statistical SLU system, we use a linear-chain CRF model; a model that assigns a joint probability distribution over labels which are conditional on the input sequences, where the distribution respects the independent relations encoded in a graph [6]. The comparison between generative model and discriminative model is shown in Fig.2.

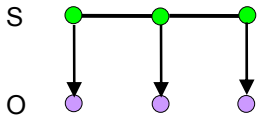
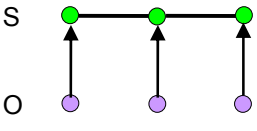
Generative Model	Discriminative Model
P(O,S)	P(S|O)
Example: HMM	Example: MaxEnt, MEMM, CRF
	
Learning=finding likelihood of observation sequence given state sequence	Learning=finding difference between state sequences given observation sequence
Tagging=finding most likely state sequence having generated given observation sequence	Tagging=finding most likely state sequence mapped from given observation sequence

Fig.2 generative model vs. discriminative model

3. Linear-chain CRF Conditional Random Field Model

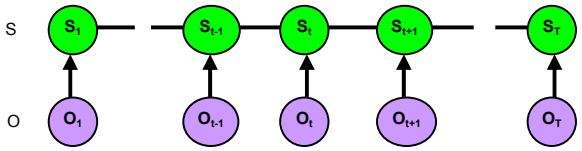


Fig.3 the structure of a linear-chain CRF

CRFs define conditional probability distributions P(Y|X) of label sequences given input sequences. We assume that the random variable sequences X and Y have the same length shown in Fig.3, and use x = x_1, …, x_n and y = y_1, …, y_n for the generic input sequence and label sequence, respectively.

A CRF on (X, Y) is specified by a vector f of local features and a corresponding weight vector λ. Each local feature is either a state feature s(y, x, i) or a transition feature t(y, y', x, i), where y, y' are labels, x an input sequence, and *i* an input position.

Typically, features depend on the inputs around the given position, although they may also depend on global properties of the input, or be non-zero only at some positions, for instance features that pick out the first or last labels.

The CRF's global feature vector for input sequence x and label sequence y is given by

$$F(y,x)=\sum_{i=1}^{T} f(y_i,x_i)$$

where i ranges over input positions. The conditional probability distribution defined by the CRF is then

$$P_\lambda(Y\mid X)=\frac{\exp(\sum_k \lambda_k\cdot F_k(Y,X))}{Z_\lambda(X)} \quad (3.1)$$

where $Z_\lambda(X)=\sum_Y\sum_k \lambda_k\cdot F_k(Y,X)$ is a normalization factor.

Any positive conditional distribution P(Y |X) that obeys the *Markov property*

$$p(Y_i\mid\{Y_j\}_{j\neq i},\mid X)=p(Y_i\mid Y_{i-1},Y_{i+1},X)$$

can be written in the form (3.1) for appropriate choice of feature functions and weight vector.

The most probable label sequence for input sequence
x is

$$\hat{y}=\arg\max_y p_\lambda(y\mid x)=\arg\max \lambda\cdot F(y,x)$$

because $Z_\lambda(X)$ does not depend on y. $F(y,x)$ decomposes into a sum of terms for consecutive pairs of labels, so the most likely y can be found with the Viterbi algorithm.

We train a CRF by maximizing the log-likelihood of a given training set $T=\{(x_k,y_k)\}_{k=1}^N$,

$$L_\lambda=\sum_k \log p_\lambda(y_k\mid x_k)$$
$$=\sum_k[\lambda\cdot F(y_k,x_k)-\log Z_\lambda(x_k)]$$

To perform this optimization, we seek the zero of the gradient

$$\nabla L_\lambda=\sum_k F(y_k,x_k)-E_{p_\lambda(Y|x_k)}F(Y,x_k) \quad (3.2)$$

To avoid overfitting, we penalize the likelihood with a spherical Gaussian weight prior

$$L'_\lambda=$$

$$\sum_k[\lambda\cdot F(y_k,x_k)-\log Z_\lambda(x_k)]-\frac{\|\lambda^2\|}{2\sigma^2}+const$$

with gradient

$$\nabla L'_\lambda=\sum_k[F(y_k,x_k)-E_{p_\lambda(Y|x_k)}F(Y,x_k)]-\frac{\lambda}{\sigma^2}$$

Iterative scaling algorithms [10] or limited-memory quasi-Newton (L-BFGS) is used to solve the parameter λ of (3.2).

4. Experiment

The test corpora contain 426 representative sentences(include open set 213 sentences) related to 30 city names. Query subjects include train number, ticket price, station name passed through, the length of a route, departure time, arrival time, stop time. These sentences are input to a speech recognizer. Speech recognizer output is the input of our SLU system which tags the semantic of text strings. The training corpora are collected from two campus BBS "ticket information". The training corpora contain 9956 sentences.

The training procedure is as follows:

Step 1: To do word segmentation and label part-of-speech tag to the sentences of the training corpora.

Step 2: To label the semantic tag of every word token of every sentence by manual labor.

Step 3: To design the template of the CRF model.

Step 4: To train the model parameter.

The test procedure is as follows

Step 1: To do word segmentation and label part-of-speech tag to the sentences of the test sentence.

Step 2: To decode the semantic sequence using the model trained in the training stage.

For saving the labor of step2 in the training stage, a portion of training corpora (noted as part 1) are first labeled by manual labor, then a model is trained on a part 1. Some of the rest training corpora (part 2) may be tagged by the model of part 1. The result tagged may be revised by manual labor. The result of part 1 and part 2 are merged into new corpora. A new model is trained on the new corpora. Such and such, finally, all of sentences in training corpora are tagged.

The average character number of sentence in test corpora and training corpora is about 10. The average number of semantic class or chunk is about 3.2. The test result is shown in Tab4.1. The correct rate of semantic unit is the correct rate of each concept or intension related to the specific domain. The correct

rate of whole sentence is the correct rate of all semantic units in the sentence.

Source	Number of sentence	Number of Semantic class	Correct rate of semantic unit	Correct rate of sentence
Closed	213	621	88%	76%
Open	213	610	86%	72%

Tab4.1 the result on the closed set and the open set

5. Conclusion and Future work

In this paper, we propose a method to Chinese spoken language understanding task using conditional random field model. SLU for spoken dialog system, especially the slot-filling problem is solved as a sequential labeling problem. The experimental results show that this approach has good performance and is feasible for the restricted domain oriented Chinese spoken language understanding.

However statistical SLU such as linear-chain CRF have good performance on the SLU task, there are a lot of semantic tag labeled firstly by manual labor. Semi-supervised learning is vital to the statistical method. From linear-chain CRF model itself, it only models local dependency feature. But in SLU there are a lot of long-distance dependency features to be modeled. Future work will focus on decreasing the labor force on labeling training corpora and exploiting the model covering local and long-distance dependency features.

Acknowledgements

This work is supported by “National Basic Research Program of China - 2006CB303105 and 2004CB318110”, NSF 60673109, Beijing Natural Science Foundation (No.4082025), Doctoral Foundation of China (No.20070004037) and Foundation of Beijing Jiaotong University (2006XM010). The methods used in the experiments have been constructed using CRF++ 5.0 Tools [13].

References

[1] Y.Y. Wang, Li Deng, A. Acero, “Spoken language understanding: an introduction to the statistical framework”, *Journal*, IEEE Signal Processing Magazine, USA, vol. 22 (5), 2005, pp. 16–31.

[2] S. Miller, R. Bobrow, R. Ingria, and R. Schwartz, “Hidden understanding models of natural language”, in Proc. 31st Ann. Meeting Association for Computational Linguistics, New Mexico State University, 1994, pp. 25–32.

[3] S. Della Pietra, M. Epstein, S. Roukos, and T. Ward, “Fertility models for statistical natural language understanding,” in Proc. 35th Ann. Meeting Association for Computational Linguistics, Madrid, Spain, 1997, pp. 168–173.

[4] Y. He, S. Young, “Semantic processing using the hidden vector state model”, *Journal*, Computer Speech & Language vol.19 (1), Elsevier, 2005, pp. 85–106.

[5] Y.-Y. Wang and A. Acero, “Combination of CFG and N-gram modeling in semantic grammar learning,” in Proc. Euro speech 2003, Geneva, Switzerland, 2003, pp. 2809–2812.

[6] J. Lafferty, A. McCallum, F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”, *Conference*, In: Proceedings of the International Conference on Machine Learning (ICML), USA, 2001, pp. 282–289.

[7] B. Taskar, C. Guestrin, D. Koller, “Max-margin markov networks”, *Conference*, In: Proceedings of the Advances in Neural Information Processing Systems (NIPS) USA, 2003.

[8] I. Tsochantaridis, T. Hofmann, T. Joachims, Y.Altun, “Support vector learning for interdependent and structured output spaces”, *Conference*, In: Proceedings of the International Conference on Machine Learning (ICML) 2004.

[9] H. Daume´III, “Practical structured learning techniques for natural language processing”, Technical Report, Los Angeles, CA. 2006

[10] A.L. Berger, S.A. Della Pietra, V.J. Della Pietra, “A Maximum Entropy Approach to Natural Language Processing:”, *Journal*, computational linguistic, vol.22(1), 1996, pp.39-73.

[11] H. M. Wallach, “Conditional Random Fields: An Introduction”, Technical Report MS-CIS-04-21, Department of Computer and Information Science, University of Pennsylvania, 2004.

[12] J. Nocedal, S. J. Wright, “Numerical Optimization”, *Book*, Springer, 1999.

[13] http://crfpp.sourceforge.net/.