Journal of Southeast University (English Edition) Vol. 23, No. 3, pp. 361 – 364 Sept. 2007 ISSN 1003—7985

Semantic role labeling based on conditional random fields

Yu Jiangde[1,2] Fan Xiaozhong[1] Pang Wenbo[1] Yu Zhengtao[3]

([1] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

([2] School of Computer and Information Engineering, Anyang Normal University, Anyang 455000, China)

([3] School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650051, China)

Abstract: Due to the fact that semantic role labeling (SRL) is very necessary for deep natural language processing, a method based on conditional random fields (CRFs) is proposed for the SRL task. This method takes shallow syntactic parsing as the foundation, phrases or named entities as the labeled units, and the CRFs model is trained to label the predicates' semantic roles in a sentence. The key of the method is parameter estimation and feature selection for the CRFs model. The L-BFGS algorithm was employed for parameter estimation, and three category features: features based on sentence constituents, features based on predicate, and predicate-constituent features as a set of features for the model were selected. Evaluation on the datasets of CoNLL-2005 SRL shared task shows that the method can obtain better performance than the maximum entropy model, and can achieve 80.43% precision and 63.55% recall for semantic role labeling.

Key words: semantic role labeling; conditional random fields; parameter estimation; feature selection

In recent years, there has been increasing interest in semantic role labeling (SRL). A semantic role is the relationship between a syntactic constituent (verb argument) and a predicate in a sentence. It identifies the role of a verbal argument in the event it is expressed by the verb: an agent, a patient, an instrument, etc. and also adjuncts such as location, time, manner, cause, etc. So, the semantic role is the role given by the verb to its arguments. The SRL task consists of analyzing and recognizing the arguments of the verbs and determining the role they play in a sentence.

For deep natural language processing (NLP), the process of fine-grain semantic role labeling is one of the prominent steps, which provides semantic relationships between constituents. The sense relationships between constituents are the core meaning of a sentence. Recently, more and more natural language processing (NLP) applications, including information extraction (IE), question answering (QA), and semantic dialogue systems are expecting support from semantic role labeling.

Some semantically-annotated corpus, such as PropBank[1] and FrameNet[2], have been manually built in English. The PropBank defines six main arguments: Arg0 to Arg5. For example, Arg0 is the agent, and Arg1 is the patient, etc. ArgM- may indicate adjunct arguments, such as location and time. FrameNet is based on the theory of frame semantics. A frame represents a scenario in terms of the interaction of its participants, and these participants play certain roles. Verbs and nouns can be used to identify a frame and the annotated sentences in each frame show the possible semantic roles for a given target word. Gildea and Jurafsky[3] were the first to apply a statistical method to the FrameNet data. They used a linear interpolation method and extracted features from a parse tree to identify and classify the constituents in the FrameNet with syntactic parsing results. Most of the following works have focused on feature engineering[4] and machine learning models[5-7].

In this paper, we apply conditional random fields (CRFs) to the task of SRL. CRFs[8] are undirected graphical models which define a conditional distribution over labels given an observation. These models allow for the use of very large sets of arbitrary, overlapping and non-independent features.

1 Conditional Random Fields

CRFs are defined as follows. Let $O = \{o_1, o_2, \ldots, o_T\}$ denote some observed input data sequences such as a sequence of phrases or named entities in training data. Let $S = \{s_1, s_2, \ldots, s_T\}$ be a set of finite state machine (FSM) states, each of which is associated with a label (such as Arg0, Arg1, ArgM-LOC). CRFs define the conditional probability of a state sequence given an input sequence O as

Received 2007-05-18.

Foundation items: The National Natural Science Foundation of China (No. 60663004), the Ph. D. Programs Foundation of Ministry of Education of China (No. 20050007023).

Biographies: Yu Jiangde(1971—), male, graduate, lecturer, jangder@bit.edu.cn; Fan Xiaozhong(1948—), male, professor, fxz@bit.edu.cn.

$$p(S \mid O) = \frac{1}{Z_O}\exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(s_{t-1}, s_t, o, t)\right) \quad (1)$$

where Z_O is a normalization factor over all candidate paths; $f_k(s_{t-1}, s_t, o, t)$ is an arbitrary feature function over its arguments; and λ_k is a learned weight for each feature function. Given such a model as defined in Eq. (1), the most probable labeling sequence for an input O,

$$S^* = \arg\max_S P(S \mid O) \quad (2)$$

can be efficiently calculated by dynamic programming using the Viterbi algorithm. Some previous studies show that two key problems in the application of CRFs are parameter estimation and feature selection.

1.1 Parameter estimation

Given the parametric form of a CRF in Eq. (1), fitting empirical distribution involves identifying the values of parameters λ_k which can be estimated by maximum likelihood, i. e. maximizing the log-likelihood L_Δ—maximizing the conditional probability of a set of label sequences, each given their corresponding input sequence. The log-likelihood of training set $\{(O_i, S_i): i = 1, 2, \ldots, N\}$ is written as

$$L_\Delta = \sum_i P_\Delta(S_i \mid O_i) = \sum_i \left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(s_{t-1}, s_t, o, t) - \log Z_{O_i}\right) \quad (3)$$

To maximize L_Δ, we have to maximize the differences among the correct paths and those of all other candidates. Lafferty et al.[8] introduced an iterative scaling algorithm for Eq. (3) L_Δ and reported that it was exceedingly slow. Several researchers have implemented gradient ascendant methods, but naïve implementations are also very slow, because various λ_k parameters interact with each other. Increasing one parameter may require compensating changes in others. Sha and Pereira[9] used the limited memory quasi-Newton (L-BFGS)[10], which is shown to be several orders of magnitude faster than iterative scaling. L-BFGS can simply be treated as a black-box optimization procedure, requiring only that the value and first-derivative of the function to be optimized be provided. In this paper, we employ L-BFGS to estimate the parameters for the CRF model.

1.2 Feature selection and feature induction

To analyze the contribution of different kinds of features, we divide the features into three categories: features based on sentence constituents, features based on predicate and predicate-constituent features. Features based on sentence constituents are these features relating to sentence constituents. The features based on sentence constituents we use are summarized in Tab. 1.

Tab. 1 List of features based on sentence constituents

Number	Feature name	Description
F_1	PhraseType	Syntactic category of the constituent
F_2	NEType	Type of NE in the constituent
F_3	HeadWord	Head word of the constituent
F_4	POS	Part of speech of the constituent
F_5	FirstWord	First word of the constituent
F_6	LastWord	Last word of the constituent
F_7	Prepositions	Preposition of the constituent
F_8	PreviousUnit	Previous semantic label of the constituent
F_9	NextUnit	Next semantic label of the constituent
F_{10}	WordNumber	Number of words in the constituent

Features based on the predicate are those features relating to the predicate of a sentence. The features we use are summarized in Tab. 2.

Tab. 2 List of features based on predicate

Number	Feature name	Description
F_{11}	PredicatePOS	Part of speech of the predicate
F_{12}	PredicatePostion	Position of the predicate
F_{13}	PredicateVoice	Grammatical voice of the predicate
F_{14}	PredicateSense	Sense of the predicate

Predicate-constituent features are those features which denote the relationship between the predicate and the constituent. The predicate-constituent features we use are summarized in Tab. 3.

Tab. 3 List of predicate-constituent features

Number	Feature name	Description
F_{15}	Path	Parse tree path from the predicate to the constituent. Fig. 1 is an example of path.
F_{16}	PathLength	The nodes number on the parse tree path.
F_{17}	Position	The relative position of the constituent and the predicate, before or after.

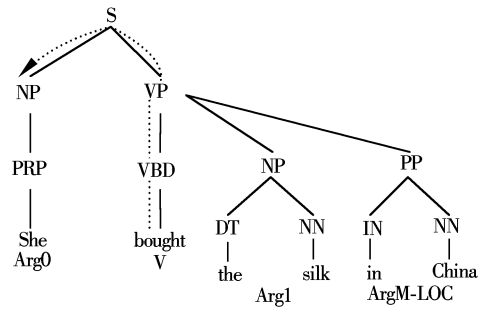

Fig. 1 In this example, the path from the predicate "bought" to the constituent "she" can be represented as VBD ↑ VP ↑ S ↓ NP

2 Using CRFs for Semantic Role Labeling

We can easily cast the SRL task as a sequence labeling problem, and apply CRFs to the task. To reduce the length of the sequence, we first transform the data from the original word-based format to a phrase-based format via shallow parsing. In this representation, the basic token is a phrase or named entities, and

its spelling and part-of-speech (POS) tags are replaced by those of its head word. Now, the SRL task can be attributed to such a computational problem: First, transform the parsing trees into the uniform syntactic representation. Secondly, the object verb and its arguments boundaries are determined. Thirdly, label the suitable semantic roles for different phrases or named entities using CRF, based on the role information and features described in the corresponding feature functions.

3 Experiments

To test the performance of the semantic role labeling based on CRFs, we trained the CRF model on training datasets. We perform different experiments under different features and different numbers of training sentences are used. For comparison we use the maximum entropy classifier to the SRL task as a baseline.

3.1 Experiment datasets and evaluation metrics

In order to build the SRL system based on CRFs, we use the benchmark corpus provided by CoNLL-2005 SRL shared task as training and test sets. The datasets consist of sections of the Wall Street Journal (WSJ) part of the Penn TreeBank, with information on predicate-argument structures extracted from the PropBank corpus. We follow the standard partition used in syntactic parsing: sections 02 to 21 for training, and section 23 for testing. The system is evaluated with respect to precision (P), recall (R), and F_1 of the predicted arguments, $F_1 = 2PR/(P + R)$.

3.2 Experimental results

3.2.1 Experiment of combined features

In any corpus-based approach a tuning process is needed in order to obtain a set of features that maximizes the results. In order to do this, we have used a k-fold cross validation evaluation method with $K = 5$. In this kind of method the training dataset is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other $k - 1$ subsets are put together to form a training set. Then the average error across all k trials is computed. The advantage of this method is that is matters less how the data gets divided. Every data point is in a test set exactly once, and is in a training set $k - 1$ times. Results regarding this tuning procedure are shown in Tab. 4. These results show that different features sets have different performances.

Tab. 4 Experimental results using different features %

Features	P	R	F_1
A random set of ten features	61.04	56.79	58.84
F_1, F_3	58.18	45.28	50.93
F_2, F_3, F_4, F_{15}	60.67	52.39	56.23
$F_1, F_3, F_4, F_7, F_{12}, F_{15}, F_{17}$	64.23	55.98	59.82
$F_2, F_3, F_5, F_6, F_8, F_{10}, F_{12}, F_{13}, F_{15}, F_{17}$	69.38	60.17	64.45
$F_1, F_2, F_3, F_4, F_7, F_8, F_{10}, F_{12}, F_{13}, F_{16}$	72.78	62.18	67.06
$F_2, F_3, F_5, F_6, F_9, F_{10}, F_{12}, F_{13}, F_{15}, F_{16}$	76.06	63.58	69.26
$F_1, F_2, F_3, F_4, F_7, F_8, F_{11}, F_{12}, F_{14}, F_{15}, F_{17}$	80.43	63.55	71.00
$F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_{10}, F_{11}, F_{12}, F_{13}, F_{14}, F_{15}, F_{16}, F_{17}$	80.26	61.33	69.53
All features	78.39	60.69	68.41

3.2.2 Experiment of different numbers of training sentences

In general, the SRL system performance varies for different numbers of training sentences used. Thus, several sets of experiments are performed, each using a different number of training sentences. As shown in Fig. 2, labeling performance steadily improves as the number of training sentences is increased.

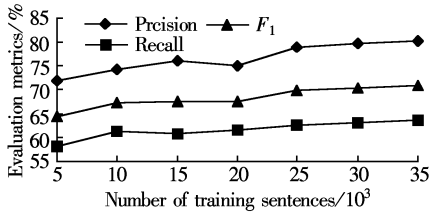

Fig. 2 Performance of the method for different sizes of training set

3.2.3 Experimental results for semantic role labeling

Tab. 5 shows detailed results of different kinds of roles using CRFs and ME on the same testing set.

Tab. 5 Experimental results for SRL using CRFs and ME %

Semantic roles	CRFs			ME		
	P	R	F_1	P	R	F_1
Arg0	86.47	81.26	83.78	85.17	82.87	84.00
Arg1	81.90	75.82	78.74	82.23	71.88	76.71
Arg2	72.36	60.18	65.71	73.69	62.81	67.82
Arg3	82.50	40.69	54.50	80.14	42.77	55.77
Arg4	86.77	65.80	74.84	82.98	61.99	70.97
Arg5	97.66	37.52	54.21	89.16	42.67	57.72
ArgM-ADV	60.98	41.66	49.52	61.02	43.78	50.99
ArgM-CAU	62.55	35.89	45.61	56.15	45.89	50.50
ArgM-DIR	67.01	34.56	45.60	60.18	44.56	51.21
ArgM-DIS	69.34	77.24	73.08	67.02	71.93	69.39
ArgM-EXT	76.39	56.02	64.64	71.98	52.47	60.70
ArgM-LOC	75.89	59.34	66.60	77.89	62.06	69.08
ArgM-MNR	66.19	50.29	57.15	64.37	51.95	57.50
ArgM-MOD	99.03	93.56	96.22	84.29	79.78	81.97
ArgM-NEG	95.97	98.16	97.05	78.06	89.23	83.27
ArgM-PRD	96.77	87.34	91.81	84.04	90.00	86.92
ArgM-PRP	87.59	79.00	83.07	83.98	69.00	75.76
ArgM-TMP	82.29	69.50	75.35	80.45	71.03	75.42
Overall	80.43	63.55	71.00	75.71	63.15	68.86

From Tab. 5, it can be found that the method based on CRFs obtains better performance than the maximum entropy model, both precision and recall of the method are raised: for the baseline from 75.71% to 80.43% for precision and from 63.15% to 63.55% for recall.

4 Conclusion and Future Work

This paper is dedicated to the problem of semantic role labeling using CRFs. Our method takes shallow syntactic parsing as the foundation, and takes phrases or named entities as the labeled units. Experimental results show that the method obtains better performance than the maximum entropy model. The main current limitation of CRFs is the slow convergence of the training algorithm. In future work, we plan to do more regarding the training algorithm.

References

[1] Palmer M, Gildea D, Kingsbury P. The proposition bank: an annotated corpus of semantic roles [J]. *Computational Linguistics*, 2005, **31**(1): 71 − 105.

[2] Baker C F, Fillmore C J, Lowe J B. The Berkeley FrameNet project [C]//*Proc of ACL &COLING*-1998. Montreal, Canada, 1998: 86 − 90.

[3] Gildea D, Jurafsky D. Automatic labeling of semantic roles [J]. *Computational Linguistics*, 2002, **28**(3): 245 − 288.

[4] Liu Huaijun, Che Wanxiang, Liu Ting. Feature engineering for Chinese semantic role labeling [J]. *Journal of Chinese Information Processing*, 2007, **21**(1): 79 − 84. (in Chinese)

[5] Pradhan S, Hacioglu K, Krugler V, et al. Support vector learning for semantic argument classification [J]. *Machine Learning Journal*, 2005, **60**(3): 11 − 39.

[6] Liu Ting, Che Wanxiang, Li Sheng, et al. Semantic role labeling system using maximum entropy classifier [C]//*Proc of CoNLL*-2005. Ann Arbor, Michigan, 2005: 189 − 192.

[7] Che Wanxiang, Zhang Min, Liu Ting, et al. A hybrid convolution tree kernel for semantic role labeling [C]//*Proc of ACL*. Sydney, Australia, 2006: 73 − 80.

[8] Lafferty J, Pereira F, McCallum A. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]//*Proc of the* 18*th International Conference on Machine Learning*. San Francisco, 2001: 282 − 289.

[9] Sha F, Pereira F. Shallow parsing with conditional random fields [C]//*Proc of Human Language Technology of NAACL*. Edmonton, Canada, 2003: 213 − 220.

[10] Byrd R H, Nocedal J, Schnabel R B. Representations of quasi-Newton matrices and their use in limited memory methods [J]. *Mathematical Programming*, 1994, **63**(2): 129 − 156.

基于条件随机场的语义角色标注

于江德[1,2] 樊孝忠[1] 庞文博[1] 余正涛[3]

([1] 北京理工大学计算机科学技术学院,北京 100081)

([2] 安阳师范学院计算机与信息工程学院,安阳 455000)

([3] 昆明理工大学信息工程与自动化学院,昆明 650051)

摘要:由于语义角色标注对深层次的自然语言处理非常必要,提出了一种基于条件随机场的语义角色标注方法. 该方法以浅层句法分析为基础,把短语或命名实体作为标注的基本单元,将条件随机场模型用于句子中谓词的语义角色标注. 该方法的关键在于模型的参数估计和特征选择. 具体应用中采用L-BFGS 算法学习模型参数,并选择基于句法成分的、基于谓词的、句法成分−谓词关系三类特征作为模型特征集. 在 CoNLL-2005 评测任务所提供的数据集上的实验结果表明:基于条件随机场的方法比基于最大熵模型的方法性能更好. 该方法在语义角色标注任务上获得了 80.43% 的准确率和 63.55% 的召回率.

关键词:语义角色标注;条件随机场;参数估计;特征选择

中图分类号:TP391