Chinese Dependency Parsing Based on An Improved Model of MST

Sun Qi[1, a], Xiang Yang[2,b] and Tu Xiao[3]

[1]Tongji University,China

[2]Tongji University,China

[3SS]Shenji Information System Engineering Co., Ltd.

[A]1556566101@qq.com, [b]shxiangyang@tongji.edu.cn

Keywords: Dependency parsing; Maximum spanning tree; Condition random fields

Abstract. In this paper, a Chinese dependency parsing method was presented based on improved Maximum Spanning Tree algorithm. Within this method, Conditional Random Field (CRF) is adopted to establish sequence labeling model. Recognizing POS of head node is employed to modify the weights of directed edges in the MST model. Comparative experiments on CoNLL 2009 data set show that the new method shows better performance than current Chinese dependency methods, with precision reaching to 85.45%.

Introduction

Dependency parsing is one of the most important tasks of natural language processing（NLP）， it has a wide range of applications on information extraction, machine learning, question answering system. It aims at converting an input sequence into a syntactic dependency tree through some dependency grammar.

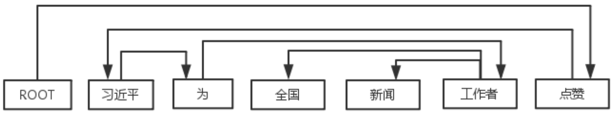

Fig. 1 – An example of dependency tree

There are two main methods of Data-driven dependency analysis, one is formula syntax analysis based on transfer and the other is based on maximum spanning tree algorithm. The transfer-based approach decomposes a dependency tree into a sequence of actions. The most appropriate transfer direction of the current state of the syntax was selected through appropriate classification, which is the next state of the sentence. The main differences among these transfer based parsing models lie in action, namely transfer strategy and the machine learning model what is used to estimate the possibility of these action, such as support machine model (SVM). Covington[1], Yamada[2] both employed this method to analyze the Chinese and English sentences. In the analysis process, SVM is used in order to eliminate syntactical ambiguity. Then the uncertainty analysis process becomes a deterministic process. Nivre[3] also adopted the method based on transfer. But their data structure and action is different. This transfer-based approach, being a greedy algorithm, can use the existing transfer state effectively. However， with the character of hard to rollback, it also brings the problem of the contagion of mistakes. It is often difficult to obtain the global optional results.

The graph-based approach based on graph treats words in a sentence as vertexes of a graph and the dependency arcs between words as the edges of a graph. The methods of machine learning are put to use to obtain the dependency probability which is the weights of directed edge. Dependency analysis is the process of finding maximum spanning tree in the directed graph. This method takes all possible dependencies in the sentence into account; therefore it is the global optional results. The research and application of graph-based approach have made great progress. Eisner presented an algorithm based on span[5], the computing complex of the search process is O(n3). McDonald proposed the first-order dependency analysis model based on graph for the first time in 2005[6].He assumes that the dependency arcs is independent and have on effect on each other. And the score of the dependency parsing tree is the sum of the score of the arcs. McDonald then proposed the second-order dependence analysis model based on graph in 2006[7]. The model allows the dependence between the adjacent brothers arcs and they are scored as a whole. Koo further proposed the third-order model which the minimum independent substructure includes three dependent arc[8]. He put forward two different third-order dependent substructure, the parent-brothers-son structure and the three-brothers-son structure. Considering the high-order subtrees enhanced the model rationality, however it increased the difficulty of decoding at the same time[8]. In the study of Chinese dependency paring, Zhou proposed the idea of chunking[9]. First group the sentence, and then separate analysis is given to some more complex chunk according to certain pre-specified rules to improve accuracy. These studies have constructive achievements and open up some new thought for dependency parsing. Zhou integrated graph-based method and transfer-based method[10]. Ma Ji integrated multiple shift-reduced model by changing the distribution of training data[11]. Nivre and McDonald proposed a combination of the two models based on their complementary relationship, introducing the parsing result of a model as a guide attribute into another model[12]. When there are little difference between the parsing accuracy of basic model, the combination model significantly improved the dependency accuracy. This paper also use the system integration method. We use the MST model as the base model. First of all, we employ the arc filters to decrease unnecessary computation and reduce the complexity of the algorithm. And then we use the sequence labeling and CRF model to modify the weights of directed edges. At last we apply the improved Chu-Liu-Edmonds algorithm to generate the final dependency trees.

Definition of the modified model

Arc Filters. Dependency parsing is to get the parsing tree that is scored highest through a scoring function which is a slow process of exhaustive search. It mainly consists of two steps include parsing and arc scoring. The general time complexity of parsing process is $O(n^3)$ while some improved method decrease the time costs to $O(n^2)$. However all these methods requires to connect the every arc to get a result tree, which always brings huge amount of calculation.

So we put forward a method of arc filtering to reduce the arc score calculation which has low probability. It is the preprocessing of dependency parsing. The dependency gragh initially includes all arcs in the gragh .That is to say, the number of arcs reaches 2^n for a sentence with n words. An arc filter which can removes implausible from the complete dependency is a supervised classifer. We filter arcs in linear time when the single node and their contexts are the main consideration. For each node, the classifers are separate include:

(1) It is a head of other node or not.

(2) Its head is on the left or right

(3) The distance between the node and its head is with 5 nodes or not.

(4) The distance between the node and its head is 1 or not.

(5) It is the root of the tree or not.

So for example, if a word is not a head of other node, then all arcs with the given word as head need to filter. If a word have a head on the left, the all arcs with the given word as child node and on the right can be pruned. In order not to reduce the accuracy , we only filter those very rarely happens .

Sequence labeling. According to the theory of dependency grammar, for a given node , if the dependency direction and position of its head word is known, the search space of its head word can be reduced. In the study of Chinese dependency parsing, Liu introduced the concept of domination[13]. It represents the word's dominant power to its child node. Ji Feng proposed a method of dependency parsing based on sequence labeling[14].In this paper, we improve this model to determine the dependency direction and distance.

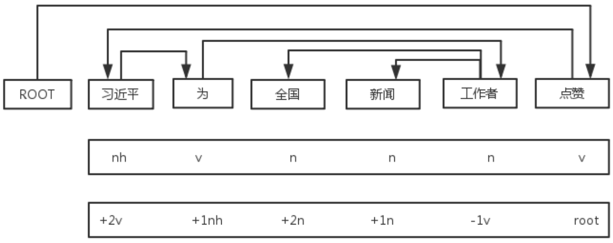

Fig. 2– Sequence labeling

We define the category labels as p=+/-dPOS, p is the POS of head node, "+/-" is the direction, "+" denotes that the head node lie in the right of the node and "-" denotes that the head node locate in the left of the node. And "d" denotes the distance between the node and its head node. When calculating the distance just consider the node what has the same POS as the head node has. For example, as is shown in figure 2. The second line contains the POS of every word. And the third line is the category labels. " by" is dependent on " Xi Jinping", so the label of " by" is "+1nh". And " Xi Jinping" is dependent on " Click a like"; so the label of " Xi Jinping" is "+2v". For a given sentence $w=w_1w_2...w_n$, it is our task to recognize the label sequence $p=p_1p_2...p_n$.

We use the CRF model to get the head node POS and use different training sets for training to reduce the POS space limitations caused by training corpus. So the validity and accuracy can be improved effectively.

Formal definition for the model and decoding algorithm

Formal definition for the model. Maximum Spanning Tree.

According to the previous definitions, $x=x_1x_2\dots x_n$ denotes the sentence to be parsed , y denotes the candidate set of dependency trees. $x(i,j)\in y$ denotes that x_j is the head node of x_i. $1\le i,j\le|x|$ and $i\ne j$. The task of dependency parsing is to find the dependent relationships between any two word. $\delta(i,j)$ denotes the dependency weights between w_i and w_j. The aim of dependency parsing based on MST algorithm is to select a spanning tree y' that satisfies dependency grammar and gets the highest score from all the candidate subtrees. y is the set of candidate dependency trees.

$$y' = \arg\max_{y} S(x,y) = \arg\max_{x} \prod_{(i,j)\in y} \delta(i,j) \tag{1}$$

Conditional Random Fields

In this paper, we use the CRF to achieve a POS labeling model, which is demonstrated more accurate than ME or perceptron model. CRF belongs to log-linear probability model. For a given input sentence, the conditional probability of a POS sequence "t" is

$$\mathrm{P}(t \mid x) = \frac{\exp(Score_{pos}（\mathrm{x}，\mathrm{t}))}{\sum_{t'} \exp(Score_{pos}(x,t'))} \tag{2}$$

$$\mathrm{Score}_{pos}(x,t) = w_{pos} \bullet f_{pos}(x,t) \tag{3}$$

$$= \sum_{1 \le i \le n} Score_{pu}(x,i,t_i) + Socre_{pb}(x,i,t_{i-1},t_i)$$

$$Score_{pu}(x,i,t_i) = w_{pu} \bullet f_{pu}(x,i,t_i) \tag{4}$$

$$\mathrm{Score}_{pb}(x,i,t_{i-1},t_i) = w_{pb} \bullet f_{pb}(x,i,t_{i-1},t_i) \tag{5}$$

“Score” denotes the POS score, “fpos(x,t)” denotes the POS feature vector, “wpos” denotes the weight vector of POS feature. We use two kinds of POS features: POS unigram features as “fpu(x,i,ti)” and POS bigram features as “fpb(x,i,ti-1,ti)” ,which is shown in table 1.

Table 1 – POS features

POS Unigram Features: $f_{pu}(x,i,t_i)$
$t_iw_i,t_iw_{i-1},t_iw_{i-2},t_iw_ic_{i-1,-1},t_iw_ic_{i+1,0},$ $t_ic_{i,0},t_ic_{i,-1},$
POS Bigram Features: $f_{pb}(x,i,t_{i-1},t_i)$
t_it_{i-1}

“$c_{i,k}$” denotes the w_i’s kth character, a Chinese symbol denotes a character, “$c_{i,0}$”denotes the w_i’s 0th character, “$c_{i,-1}$” denotes the w_i’s last character. Decode for the input sentence, using the classical Viterbi algorithm.

Firstly, operate the MST model on each input sentence to get every possible dependency arc. Then adopt the POS labeling model to modify the weights of directed edge. Ultimately obtain the dependence tree by Chu-Liu-Edmonds search algorithm, as is shown on Figure 3.

Searching Algorithm. In this paper, we use the improved Chu-Liu-Edmonds algorithm as the search algorithm of maximum spanning tree. As we know, the Chu-Liu-Edmonds algorithm calculates every arc in the completely directed graph, which needs considerably long computational time. So we use the arc filtering as a preprocessing step for calculate the score of every arc.

The data for model training are annotated trees. For each sentence, all possible arc can be extracted and those appears or not in the annotated tree are labeled as class -1/+1. In the same way, the training data for the other filters are also generated. Our goal is to only filter those implausible arcs to not affect the accuracy of dependency. So we have a high tendency to those classifier that has high cancellation and reduce the the cost effectively for classify a true arc during learning.

In many learning package ,class-specific cost is command line parameter. The learning objective can be interpreted as reducing regularized and the weighted loss is :

$$\min_{\bar{w}} \frac{1}{2} \| \bar{w} \|^2 + C1 \sum_{y_i=1} l(\bar{w}, y_i, \bar{x}_i) + C2 \sum_{y_i=-1} l(\bar{w}, y_i, \bar{x}_i) \tag{6}$$

In this formula, l() is the loss function based on the learning method, $\overline{x_i}$ are the features and and y_i is label for the i-th training example, $\bar{w}$ is the learned weight vector, and C_1 and C_2 are the class-specific costs. When C2>>C1 you can get a higher accuracy. For an SVM, $l(\overline{w}, y_i, \overline{x_i})$ is the standard hinge loss.

After remove the implausible arcs from the complete dependency graph, we use the Chu-Liu-Edmonds algorithm introduced by McDonald to get the maximum spanning Tree. The steps are as follow:

Step 1: Calculate every possible arc's score in the directed graph by the trained model;
Step 2: Leave the highest score incoming edges for each node (not include the root node);
Step 3: Judge that if there is ring in the generated graph, if not, that is what we want. Or skip to step 4;
Step 4: Identify every node in the ring and the largest score adjacent node outside, add the arc not in the ring and remove the incoming edge in the ring. And then skip to the Step 3.

Experiment analysis

CoNLL 2009 data set is adopted in experiment. The training set include 22277 sentences and with the average sentence length of 27.43. And the testing set include 1762 sentences and the average sentence length reach to 28.16. Evaluation standard uses the UAS, namely the percentage of the correct core node in the test data set.

We adopt the CRF++ tools to train the sequence labeling model, compares the classical maximum spanning tree algorithm. Results are shown in the Table 2 and Figure 4.

Table 2– UAS results

The length of sentence	MST	Ours
<20	88.35%	89.42
20-30	86.99%	87.63%
30-40	86.59%	87.43%
ALL	85.13%	86.61%

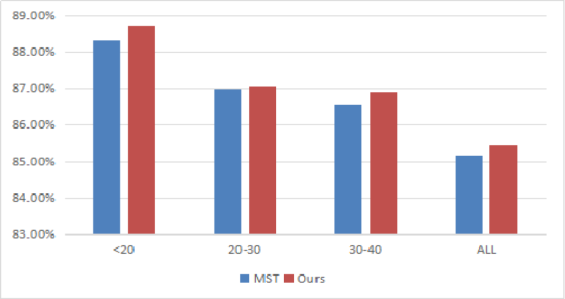

Fig. 2– Results of dependency parsing

The results show that for the each sentence length, the improved algorithm based on the maximum spanning tree can effectively improve the accuracy of dependency.

In the second experiment, we measured the filtering results. We define Reduction as the percentage of arcs filtering and coverage as the percentage of true arcs retained .We compare to the Vine parsing approach Eisner and Smith introduced. The results are shown in the Table 3.

Table 3– Performance(%) of filters on test data

Filter	Coverage	Reduction
Vine	99.62	44
Ours	99.75	53

The result show that Our arc filtering algorithm not only improve the efficiency of the filter, but also retain more true arcs. Using the filtering algorithm as a preprocessing step for dependency parsing can preserve high stabilization accuracy ,at a dramatic reduction of computational complexity.

Conclusion and future work

Chinese dependency parsing based on maximum spanning tree algorithm has gained good analytical performance. In this paper, we take advantage of the sequence labeling model to modify the weight of directed edges in the MST. Then use the improved Chu-Liu-Edmonds algorithm get the dependency results. Comparative experiments on CoNLL 2009 dataset show that the new method shows better performance than current Chinese dependency methods, with precision reaching to 85.45%.

About the direction for future study, the accuracy of sequence labeling plays an important role in the dependency parsing, and affects the correctness of parsing directly. On this basis, a further research is conducted into the accuracy of sequence labeling.

References

[1]. Michael A. Covington. A fundamental algorithm for dependency parsing[C]. Proceedings of the 39th Annual ACM Southeast Conference, 2001.

[2]. Yamada, H., and Matsumoto, Y. 2003. Statistical dependency analysis with support vector machines. In IWPT2003, pp. 195-206.

[3]. Joakim Nivre, Mario Scholaz. Deterministic Dependency Parsing of English Text[C]. Proceedings of COLING. 2004: 64-70.

[4]. R. McDonald. Discriminative learning and spanning tree algorithms for dependency parsing. Ph.D. thesis, University of Pennsylvania, 2006,10-18.

[5]. Joakim Nivre, Mario Scholz. Deterministic Dependency Parsing of English Text[C]. Proceedings of COLING. 2004:64-70.

[6]. Ryan McDonald, Koby Crammer and Fernando Pereira. Online Large-Margin Training of Dependency Parsers[C]. Association for Computational Linguistics (ACL). 2005.

[7]. Ryan McDonald, Fernando Pereira. Online Learning of Approximate Dependency Parsing Algorithms[C].European Association for Computational Linguistics (EACL). 2006.

[8]. Koo, T., and Collins, M. Efficient third-order dependency parsers. In ACL2010, pp. 1-11.

[9]. Zhou Q, Huang CN. An improved approach for Chinese parsing based on local information. Journal of Software,1999.

[10]. Zhou Huiwei, Huang Degen etc, MST Parsing Algorithm and Nivre's algorithm Integrated Dependency Parser for Chinese, Advances of Computational Linguistics in China,2011.

[11]. Ma Ji, Zhu Muhua, Single-Model System Combination for Shift-Reduce Parser, Advances of Computational Linguistics in China,2011.

[12]. Nivre J, Mcdonald R T. Integrating Graph-Based and Transition-Based Dependency Parsers.[C]// ACL 2008, Proceedings of the, Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, Usa. 2008:950--958.

[13]. Liu Ting, Ma Jin-Shan, Li Sheng. Chinese Dependency Parsing Model Based on Lexical Governing Degree. Journal of Software , 2006.pp.1876-1883