基于序列标注的中文依存句法分析方法

计 峰 邱锡鹏
(复旦大学计算机科学与工程系 上海 200433)

摘 要 提出了一种基于序列标注模型的中文依存句法分析方法。该方法将依存句法分析转化成序列标注问题,利用条件随机场 CRF(Conditional Random Field)建立序列标注模型。在宾州中文树库的测试中,达得了 76.59%的依存关系准确率,句子准确率也达到了 23.5%。同时我们改进了 Viterbi算法,使得依存关系的准确率提高了近 2 个百分点,句子准确率提高了近 3.5 个百分点。

关键词 依存分析 条件随机场 Viterbi算法

A NEW CHINESE DEPENDENCY ANALYSIS METHOD BASED ON SEQUENCE LABELING MODEL

Ji Feng Qiu Xipeng
(Department of Computer Science and Engineering, Fudan University, Shanghai 200433, China)

Abstract In this paper, a new Chinese dependency analysis method based on sequence labeling model was proposed. The problem was transformed into a sequence labeling problem by utilizing conditional random field model. The test in Penn Chinese Treebank version 2.0, a small-scale corpus, has shown the result of around 72.9% dependency accuracy and around 23.5% sentence accuracy. Meanwhile, we improved the Viterbi algorithm, and the final performance can be improved about 2% on dependency accuracy and 3.5% on sentence accuracy.

Keywords Dependency analysis Conditional random field Viterbi algorithm

0 引 言

不同于短语文法,依存文法理论认为每个句子中存在一个唯一的中心词,支配着句子中其他所有的词,其他词直接或间接依赖于中心词;同时句子中除了中心词外每个词都只被一个词支配。依存文法可以使用依存句法树表示,如图 1所示(例句“第七届世界游泳锦标赛今晚在罗马开幕。”对应的依存句法树),有向弧直接连接存在直接依存关系的两个词汇,有向弧的方向从支配词指向从属词。不同于经典的依存分析方法[1-3],本文提出了一种全新的依存分析方法,通过将依存句法分析问题转化为序列标注问题,利用 CRF[4] 建模,同时通过对解码算法的改进得到了一种性能较高的依存句法分析方法。

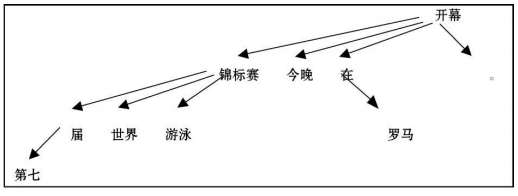

图 1

1 基于序列标注方法的依存句法分析

依存句法分析本质上可以转换为分类问题,因此依存句法分析问题实际上也是可以转化成序列标注问题。如果直接使用词作为序列标签是不合适的,因为这样会导致标签数量过多,无法建模。因此我们首先将树库转换成适合序列标注的语料。

1.1 树库转换

首先,根据依存文法理论,我们可以知道决定两个词之间的依存关系主要有二个因素:方向和距离。因此我们将类别标签定义为具有如下的形式:

$$[+|-]d\mathrm{POS}$$

其中,[+|-]表示方向,+表示支配词在句中的位置出现在从属词的后面,-表示支配词出现在从属词的前面;POS表示支配词具有的词性类别;d表示距离。

d表示的距离不是指表层距离(表层距离定义为两个词在句子中的位置之差),而是指从某个方向开始第 d 个具有相同词性为 POS的词。如图 2(例句“第七届世界游泳锦标赛今晚在罗马开幕。”转换成标注序列)中,“锦标赛”受到“开幕”的支配,两个词表层的距离为 4,而我们标签中 d的值应为 1。由于句子的中心词并不受句中其他词的支配,所以定义句子中心词的标签为 -1ROOT,相当于在句首添加了一个虚词 ROOT。

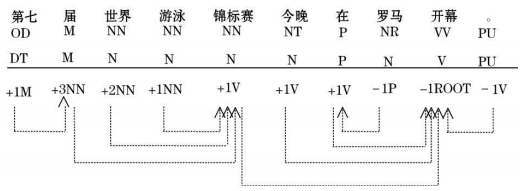

图 2

收稿日期:2008-05-06。计峰,硕士生,主研领域:自然语言处理,信息检索。

对于 POS词性的粒度, 我们采用了混合方式确定。通常在词性标注规范中, 词性可以分成两层: 一层粒度较大, 另一层粒度较小, 如图 2中横线上方的两行。所谓的混合方式是指标签中 POS部分使用了不同粒度的词性。如果依存关系中支配词不是名词, 那么 POS使用支配词粒度较大的词性; 如果是名词, 那么 POS为支配词粒度较小的词性。

图 2中第 1行表示句子中的词; 第 2行表示对应粒度较小的词性; 第 3行表示词性标注规范中对应词性上层粒度较大的词性; 横线下的一行是经过转换后的标签。虚线是根据标签可以得到对应的支配词, 从而构成一棵完整的依存句法树。

可以看出, 转换后的类别标签序列和原有的依存关系是一一对应的, 同时能够保持信息的完整性。这样定义每个词的类别标签主要有两个方面的好处:

a) 大大减少了用于序列标注的类别数量。通过在依存文法树库上的统计, 如果直接使用支配词作为类别, 那么标签的数量就是词表的大小, 而一般词表的大小都要在几万的数量级。通过我们的方法转换后的标签数量一般在 150– 220个。

b) 相对缩短了具有依存关系的两个词之间的距离。根据语言的组织习惯, 从属词通常出现在支配词的邻近周围, 表层距离较近。这样的假设使得短距离的依存关系要比长距离的依存关系具有更高的优先级, 算法会倾向于短距离的依存关系, 导致长距离依存关系很难被正确分析。而我们定义的标签使用了存在依存关系的两个词间与支配词具有相同词性的词的数量作为距离。这样定义的距离最大等于表层距离, 最小为依存距离 1, 相对缩短了两个词之间的距离。

1.2 CRF和特征选择

CRF[4]是定义在一个无向图上的指数概率模型, 其中最简单的形式是线形链式 CRF。

假设给定一个观察序列的随机变量 $X = (x_1, x_2, \cdots, x_n)$, 以及相对应标注序列的随机变量 $Y = (y_1, y_2, \cdots, y_n)$, 其中 x_i表示 X 的第 i个分量, y_i表示 x_i对应的标签。线形链式 CRF定义为这样一个条件分布:

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{n} \sum_{k=1}^{K} \lambda_k f_k (y_i, y_{i-1}, x) \right)$$

其中 $Z(x) = \sum_{y \in Y} \exp\left(\sum_{i=1}^{n} \sum_{k=1}^{K} \lambda_k f_k (y_i, y_{i-1}, x) \right)$为归一化因子; $f_k(y_i, y_{i-1}, x)$ 为特征函数, 共 K 个。特征函数 $f_k(y_i, y_{i-1}, x)$ 可以分为两大类: 一类为只与当前位置的标签相关的特征函数 $f_k(y_i, x)$, 类似于隐马尔科夫模型中观察到当前状态时的特征; 另一类为与当前和前一个位置的标签相关的特征函数 $f_k(y_i, y_{i-1}, x)$, 对应于隐马尔科夫模型中的发生状态转移时的特征。

从以上的定义中可以看出, CRF可以充分利用上下文信息作为特征, 同时能够使用复杂、非独立的特征, 从而使得 CRF模型的表达能力大大提高, 并具有很强的推理能力。同时 CRF解决了最大熵模型中存在的"Label Bias"问题。

对于序列标注问题, 我们最终需要求解一个最优的序列, 即:

$$y^* = \arg\max_{y \in Y} p(y \mid x)$$

根据线形链式 CRF的特性, 最优序列可以通过 Viterbi算法[5]求解。

线形链式 CRF模型参数的估计属于最大似然估计, 所以可以使用 EM 算法, 参数的优化可以使用 LBFGS算法。

我们通过特征模板来抽取特征。特征模板具体如下:

```
# Unigram
w0                      w0p0
p-1w0                   p-1p0w0
w0p1                    w0p0p1
p-1p0                   c-1p0
p0p1                    p0c1
p-2p-1p0p1              p-1p0p1p2
p-2p-1p0p1p2            c-2p-1p0
p-1p0p1                 c-1p0c1
p-1p1

# Bigram
c-1p0                   p0c1
```

其中 w 表示词, p 表示小类词性, c表示大类词性。下标表示相对于正在抽取特征的词的位置, 如图 2例句中若 w_0是"界", 那么 w_{-1}为"第七", p_1为 NN。

一元模板 (Unigram)定义的特征表示只与当前位置对应的标签相关的特征 $f_k(y_i, x)$; 二元模板 (Bigram)定义了前一个位置和当前位置对应的标签相关的特征 $f_k(y_i, y_{i-1}, x)$。如当前位置的词为"界", 那么一元模板 w_0定义了这样一个指示函数:

$$f_{U00} = \begin{cases} 1 & \text{如果 } w_0 = \text{"界"} \quad p_0 = \text{"+ 3}MM\text{"} \\ & \text{或 } w_0 = \text{"界"} \quad p_0 = \text{"+ 2}MM\text{"} \\ & \cdots \\ 0 & \text{其他} \end{cases}$$

而二元模板 $c_{-1}p_0$定义了以下的指示函数:

$$f_{B17} = \begin{cases} 1 & \text{如果 } c_{-1} = \text{"}DT\text{"} \quad p_0 = \text{"}M\text{"} \\ & y_{-1} = \text{"+ 1}M\text{"} \quad y_0 = \text{"+ 3}MM\text{"} \\ & \text{或 } c_{-1} = \text{"}DT\text{"} \quad p_0 = \text{"}M\text{"} \\ & y_{-1} = \text{"+ 1}M\text{"} \quad y_0 = \text{"+ 2}MM\text{"} \\ & \text{或 } \cdots \\ 0 & \text{其他} \end{cases}$$

1.3 算法改进

由于 Viterbi算法[5]在求解最优标注序列时需要计算每个词被标注为整个训练集上出现的所有标签的概率, 而在我们的标签体系中, 每个标签实际指向了依存关系中的支配词, 本身受到句子的约束。这样产生了一个问题, 即在没有限制的最优序列中可能出现超越句子范围的标签。如图 2中"届"一词的类别标签可能被标注成 + 4NN, 而从"届"的位置开始往后最多只出现了 3个词性为 NN的词, 使得 + 4NN 的标签对于"届"是没有意义的。因此我们改进了 Viterbi算法, 使其能够在求解最优标注序列时, 自动排除没有意义的标签, 最终得到一个受到句子约束的最优序列。具体的算法如下:

初始化:

$$\delta_1(i) = \pi_i b_i(w_1) \quad 1 \leqslant i \leqslant n$$

$$\varphi_1(i) = 0$$

递推:

for($t = 1$:T)

if(状态 j是有意义)

$$\delta_t(j) = \max_{1 \leqslant i \leqslant n} [\delta_{t-1}(i) a_{ij}] b_j(w_t)$$

$$\varphi_t(j) = \arg\max_{1 \leqslant i \leqslant n} [\delta_{t-1}(i) a_{ij}]$$

终止:

$$p^* = \arg\max_{1 \leqslant i \leqslant n} [\delta_T(i)]$$

回溯:

$$q^*_{大} = \varphi_{t+1}(q^*_{t+1}) \qquad t = T-1, T-2, \cdots, 1$$

2 实验结果与分析

宾州中文树库 2.0版选用了从 1994年到 1998年新华社发表的 325篇新闻, 包含 4153个句子, 大约 100 000词。由于宾州中文树库是根据短语文法建立的, 因此在进行实验前, 我们首先使用了 Penn2Malt工具将宾州中文树库 2.0版的短语文法树库转换成了依存文法树库。最终我们选择了编号是 301篇的 300篇文章作为训练集, 编号 301- 325的 25篇文章作为测试集。其中训练集包含 3800个句子, 测试集包含 353个句子。

对依存关系的评价, 我们使用了依存关系的准确率 DA (Dependency Accuracy)、句子中心词的准确率 RA (Root Accuracy)以及依存关系完全正确的句子准确率 SA (Sentence Accuracy)作为我们的评测指标。同时也评价了不同长度句子的依存关系的准确率。

我们的第一个实验是对比了原始 Viterbi解码算法和我们提出的改进算法的性能区别。表 1中的 System 1是使用了原始 Viterbi解码算法的结果, System 2使用了我们的改进算法的结果。从表 1中可以看出, 经过改进后的 Viterbi算法对于我们的任务在每项指标上的提高都是很明显的, 大部分的提高幅度都在 2%以上, 特别, 整句的准确率 (SA)提高 3.4%; 对于长度不超过 20词的句子也提高了 3%。

同时我们可以看出, 改进后的系统对于标点符号的精度并没有实质的提高。System 1的不计标点的总体准确率相比较于计算标点的总体准确率提高了 0.37%, 而 System 2中也只提高了 0.35%。我们认为标点符号对于分析长距离依存关系是非常有帮助的, 而对于标点的依存关系分析需要全局上下文信息, 但我们使用的特征只是反映了局部范围的上下文。

表 1 线性 CRF Viterbi解码算法和改进后的 Viterbi算法的性能比较

	DA					SA	RA
	< 20	< 40	< 100	Total	Total(no punc)		
System 1	0.829	0.759	0.734	0.729	0.766	0.235	0.678
System 2	0.859	0.784	0.756	0.750	0.785	0.269	0.690

表 2 与其他系统的性能比较

	DA					
	< 10	< 20	< 30	< 40	< 50	Total
LIU	0.861	0.774	0.744		0.739	0.74
Ours	0.918	0.859	0.814	0.784	0.772	0.750

同时我们列出了与 LIU Ting[2]提出的词汇支配度模型的中文依存句法分析方法的对比结果。LIU Ting的系统使用的依存句法树库是由哈工大建立的中文依存树库。该语料的规模比宾州中文树库 2.0版要大得多, 包含了 46000句取自《人民日报》的句子, 其中 40000句作为训练集, 2000句作为开发集, 4000句作为测试集。尽管这样的比较不一定合理, 不具有代表性, 但是也可以从一个侧面看出我们在一个规模小很多的语料上达到了更高的性能, 特别是对于句子长度不长的句子。

3 总结与展望

依存文法凭借其表达简洁、易于标注等特点, 逐渐成为句法分析领域的研究热点。在本文中, 我们提出了一种基于序列标注的中文依存文法分析方法。通过实验分析, 可以看到我们提出的方法在语料规模相对较小的情况下达到了比其他系统更好的性能。

但是线形 CRF的序列标注模型只能结合局部范围内的线性特征, 本身与依存句法树的结构化表示并不一致。由此也产生了一些新的问题, 如一个句子的最优标注结果中可能出现多个中心词, 特别是对于句子长度较长的句子。产生这些问题的根本原因在于句子中长距离依存关系很难被 CRF的局部特征捕获到。

然而我们的方法达到的结果给予我们很大的鼓舞。下一步我们的工作将重点解决在 CRF模型中引入依存句法分析树内结构化的高阶特征, 使模型更能表达依存关系, 以期待能够获得更大的性能提高。同时我们认为将长句分割成短句, 即先对短句分析依存关系, 再将短句组合成长句的完整依存关系, 对于长距离的依存关系分析将有很大的帮助, 因此也将成为将来的研究方向之一。

参 考 文 献

[1] Eisner J. Three new probabilistic models for dependency parsing: An exploration. In Proceedings of the COLING, Copenhagen, 1996: 340-345.

[2] Liu T, Ma J, Li S. 基于词汇支配度的汉语依存分析模型[J]. 软件学报, 2006, 17(9): 1876-1883.

[3] McDonald R, Pereira F, Ribarov K, et al. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In Proceeding of Human Language Technologies and Empirical Methods in Natural Language Processing (HLT-EMNLP), 2005.

[4] Sutton C, McCallum A. An Introduction to Conditional Random Fields for Relational Learning. In Lise Getoor and Ben Taskar, editors, Introduction to Statistical Relational Learning. MIT Press.

[5] Rabiner L R. A tutorial on hidden Markov Models and selected applications in speech recognition. In Proceedings of the IEEE 77(2), 1989: 257-286.

(上接第 113页)

建议性的对策探讨, 供同仁们参考。但由于就业管理工作是一项系统性、复杂性的工程, 在今后的工作当中, 还需不断地根据新的实情继续进行各种研究和探讨, 以适应新形势的需要。

参 考 文 献

[1] 教育部. 中国教育统计年鉴 1999[S]. 北京: 人民教育出版社, 2000: 377-340.

[2] 刘梦. 高职毕业生不愿当"蓝领"[N]. 中国教育报, 2002-02-02.

[3] 朱健. 择业难过哪道坎儿? [N]. 中国教育报, 2002-09-04.

[4] 高举邓小平理论伟大旗帜, 把建设有中国特色社会主义事业全面推向二十一世纪[R]. 北京: 人民出版社, 1997: 40.

[5] 吕福源. 高校扩招的同时要做好就业工作[N]. 中国教育报, 2000-10-19.

[6] 池忠军. 把握大学生择业观的变化, 积极务实做好就业指导[J]. 思想政治教育导刊, 2003(4).