第 33 卷 第 3 期 计 算 机 工 程 2007 年 2 月
Vol.33 No.3 Computer Engineering February 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)03—0172—02 文献标识码：A 中图分类号：TP391

基于 HMM 的汉语介词短语自动识别研究

奚建清，罗 强
(华南理工大学计算机学院，广州 510641)

摘 要：提出了一种基于隐马尔可夫模型(HMM)的介词短语界定模型，通过 HMM 的介词短语边界自动识别和依存语法错误校正 2 个处理阶段，较好地完成了对一个经过分词和词性标注的句子进行介词短语界定任务，为更进一步的句法分析工作打下良好的基础。试验结果显示：该模型的识别正确率达到了 86.5%（封闭测试）和 77.7%(开放测试)，取得了令人满意的结果。

关键词：汉语介词短语；自动识别；依存语法

Research on Automatic Identification for Chinese Prepositional Phrase Based on HMM

XI Jianqing, LUO Qiang
(College of Computer, South China University of Technology, Guangzhou 510641)

【Abstract】 This paper describes an automatic prediction model of Chinese prepositional phrase boundary location based on HMM. It consists of two stages: automatically identify the phrase boundary using statistics from treebank, then, post-tune the results with dependency grammar knowledge generated by dependency treebank. Experimental results demonstrate a high rate of success for predicting boundary location (86.5% correct rate for close testing and 77.7% for open testing).

【Key words】 Chinese prepositional phrase; Automatic identification; Dependence grammar

1 概述

给定一句经过切分和词性标注的句子，如何利用其中的词语、词性和句法特征信息，确定介词短语 (Chinese Prepositional Phrases, CPP)的位置，即哪个词出现在的左边界，用“[w”表示，哪个词出现在 CPP 的右边界，用“w]”表示，其中 w 是词语，如对于汉语句子“用/p 先进/a 典型/n 推动/v 部队/n 全面/a 建设/n 。/wp”，经过 CPP 自动界定处理后，应能得到结果：“[用/p 先进/a 典型/n] 推动/v 部队/n 全面/a 建设/n 。/wp”，这是 CPP 自动界定研究需要解决的问题。由于介词短语在现代汉语中的使用频率很高，此问题的正确解决，对于进一步进行句法分析具有重要意义，因为一旦能够较好地识别出 CPP，就可以减少错误发生的概率，提高句法分析的精度，为自然语言应用领域提供高质量的服务。

近年来，许多研究者在介词短语的自动界定方面做了一些有益的探索。在英语方面有代表性的工作包括：Eric Brill[2]的基于启发式规则的转换算法，Patrick Pantel[1]的基于语料库的无指导的学习方法，Lee Schwartz[3]提出的双语对齐消歧法和Mark McLauchlan等[4]的相似词语平滑算法等。与英语相比，汉语介词短语的自动识别则更为困难，这主要是由于CPP的层次不清晰、语法结构比较丰富的特性决定的。从近几年的情况看，我国学者在汉语基本短语(baseNP)的界定方面已经积累了相当的经验，并取得了令人满意的效果[5~10]，但遗憾的是直到目前为止对CPP的自动识别工作仍未见诸 报道。

CPP 的自动识别与 baseNP 的自动识别既有联系又存在着差别，一方面它们都具有明显的界限特征，因此在进行识别的时候可以采用类似的统计方法；但另一方面，CPP 在结构上比 baseNP 更复杂，因此仅依赖边界的分布信息势必引起更多的分析歧义。根据这个特点，我们提出了一种基于 HMM 的 CPP 自动识别模型，该模型结合了 CPP 左右边界词语的依存语法知识，有效消除了分析歧义。实验显示，该模型在封闭环境下对 CPP 自动识别的正确率达到了 86.5%、召回率达到了 84.6%，在开放环境下的正确率为 77.7%、召回率为 75.1%，获得了较为满意的结果。

2 CPP 的自动识别模型

2.1 基本统计模型设计

令 S=<W,T>为准备进行介词短语分析的句子，其中 $W = w_1 w_2 ... w_n$ 为词语序列，$T = t_1 t_2 ... t_n$ 为词性序列。设 b_0, b_1, b_2 分别对应不划分，左划分和右划分 3 种界定类型。这样介词短语的界定任务就转化成求解一个划分序列 $B = b^{(1)} ... b^{(n)} (b^{(i)} \in \{b_0, b_1, b_2\}, 1 \le i \le n)$，使得

$$B = \arg\max_{B'} P(B'|<W,T>) = \arg\max_{B'} P(<W,T>|B')P(B') \quad (1)$$

其中 B' 表示一种可能的划分序列。

假设词语间的界定类型是独立的，同时考虑到词语对词性的选择具有一定联系，则可得

$$P(<W,T>|B) \approx \prod_{i=1}^{n} P(w_i, t_i | b^{(i)}) P(t_i | b^{(i)}) \quad (2)$$

对于 $P(B')$，利用二元模型进行简化，得到

基金项目：国家“十五”科技攻关计划基金资助重点项目（A3480266）；广东省自然科学基金资助项目（B6480598）

作者简介：奚建清(1963－)，男，教授、博导，主研方向：知识管理，智能信息检索等；罗 强，博士生

收稿日期：2006-02-21 **E-mail：**qluo163@163.com

$$P(B')=P(b^{(0)})\prod_{i=1}^{n}P(b^{(i-1)}|b^{(i)}) \quad (3)$$

将式(2)和式(3)代入式(1)，即可得求得 CPP 划分的一般统计公式：

$$B=\arg\max_{B'}\prod_{i=1}^{n}P(w_i,t_i|b^{(i)})P(t_i|b^{(i)})P(b^{(i-1)}|b^{(i)}) \quad (4)$$

用 HMM 的 Viterbi 算法计算 CPP 划分的最佳路径，就可以初步实现对一个句子的介词短语界定。

2.2 基于词依存语法(DG)的错误界定自动校正

利用基本统计模型处理的结果含有比较多的分析歧义，经过与人工处理的结果比对发现错误主要表现在两个方面 ：(1)右边界存在多种界定结果(约占 80%)；(2)右边界误识或者漏识(约占 18%)，这说明基本统计模型对左边界的界定效果比较令人满意，而对右边界的界定精度却有待提高。我们以此作为错误校正的出发点，提出了一个基于依存语法的错误自动界定方法，该方法利用依存树库中的 CPP 的句法特征信息，从有限多个右边界词语中选择一个最合适的词语与左边界词语形成介词短语搭配，以降低错误界定发生的几率。

易知错误校正的关键在于寻找一个词语 w_j，使 w_j 和左边界词语 w_i 具有最大的语义关联度。这样，通过参考依存语法中关于词与词的依存关系的概念，可以将错误校正任务转化为求取 w_i 和 w_j 的最大的依存度值的问题。下面给出用依存语法表示 CPP 的两个定义：

定义 1 令 $<w_i,t_i>$ 和 $<w_j,t_j>$ 为一个介词短语结构的左边界和右边界，那么 w_i 和 w_j 具有依存关系，且 w_j 依存于 w_i，表示为 $w_i \leftarrow w_j$，其中 w_i 称为中心词。

定义 2 设 CPP 的词语序列为 $w_i...w_{i+k}$，w_i 为介词短语的左边界词，右边界词 $w_j(0<j\le n)$ 必满足条件：

$$j=\arg\max_{1\le i,j\le n}P(attachment|<w_i,t_i>,<w_j,t_j>) \quad (5)$$

其中 *attachment* 表示 w_j 是否依附于 w_i 的标志(0－不依附，1－依附)。

定义 1 给出了 CPP 左右边界词语在依存语法中的关系性质，定义 2 给出了寻找右边界的条件。但由于数据稀疏的原因，在实际的语料库中很难得到理想的计算结果，下面对式(5)做进一步近似。假设左边界词语是固定的，根据贝叶斯公式，可得

$$P(attachment|<w_i,t_i>,<w_j,t_j>)$$
$$\approx P(<w_{j-1},t_{j-1}><w_j,t_j>|b^{(j)})P(<w_i,t_i>) \quad (6)$$

用 t_j 和 t_{j-1} 近似代替上式中的 $<w_j,t_j>$ 和 $<w_{j-1},t_{j-1}>$，这样式(6)进一步化简为

$$j=\arg\max_{i<j\le i+k}P(t_{j-1},t_j|b^{(j)},dist(i,j))P(<w_i,t_i>) \quad (7)$$

其中 $dist(i,j)$ 表示 w_i 和 w_j 之间的距离。

另外，从哈工大的依存共享树库[11]中，可以提取CPP的句法特征信息，这些信息包括CPP的左右边界词语的共现频度、CPP左右边界词性的共现频度和右边界的上下文信息（目前，只考虑右边界前一个词语的语法信息，即观察窗口大小为 2）。利用这些数据，通过MLE方法进行参数估计，然后带入式(7)进行计算和比较，即可实现右边界的选择。

3 实验和讨论

实验对基于HMM的CPP自动识别模型和基于DG错误校正过程的性能进行了测试。实验数据来自哈工大共享依存树库[11]中包含的介词短语句子共 4 966 句，每个句子占 3 行，第 1 行是经过分词和词性标注的句子，第 2 行在句子中每个词及词性的前面加上序号，句子的末尾增加一个句尾标志“ <EOS> ”，由其支配全句的核心词，第 3 行是句子中词与词之间的依存关系，所有的数据均由人工标注。选取其中的 3 466 条句子为训练集，封闭测试集为取自训练集的 1 500 条句子，开放测试集为取自训练集之外的 1 500 条句子。

3.1 测试数据的评价

对于系统的整体处理性能，本文主要采用了以下两个评价指标：

(1)召回率，其计算公式为

召回率＝(正确处理的界定数目)/(训练集的界定数目)×100%

(2)识别精度，它反映了经过模型处理的界定情况中正确界定的比率。其计算公式为

识别正确率＝(正确处理的界定数目)/(模型处理的界定数目)×100%

3.2 介词短语识别结果分析

对于 CPP 自动识别模型的整体性能主要考虑以下两个处理过程的性能指标：统计模型识别（SI），基于词依存关系的错误校正（DGC），得到以下测试结果（见表 1、图 1）：

(1)基本性能测试

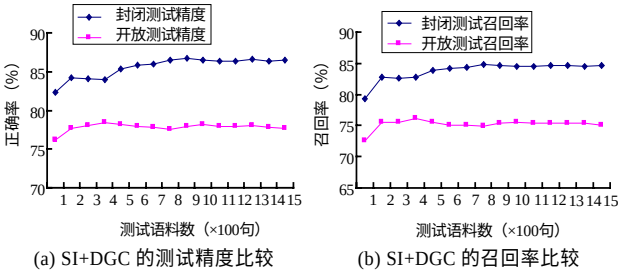
表 1 CPP 自动识别处理的基本情况

%	SI		SI+DGC	
	Precision	Recall	Precision	Recall
开放测试	37.0	93.8	77.7	75.0
封闭测试	33.5	84.5	86.5	84.6

从表 1 显示的数据来看，经过 DGC 处理后的识别精度有了较大改善，分别提高了约 41%（开放测试）和 53%（封闭测试），但召回率却有不同程度的降低，下降的幅度分别为封闭测试的 0.1%和开放测试的 19%。此外，在开放环境和封闭环境下的实验结果存在着明显差异，如在 SI 处理中开放测试的识别正确率和召回率都略高于封闭测试下的结果（分别为 3.5%和 9.3%），而在 SI+DGC 处理中封闭测试的识别正确率和召回率却高于开放测试下的结果(分别为 8.8%和 9.6%)，这说明 CPP 自动界定模型对训练语料具有一定程度的依赖性。

(2)稳定性测试

将封闭测试的 1 500 条句子和开放测试的 1 500 条句子，从第 100 条句子出发，每次增加 100 条句子，形成两个具有 15 个测试子集的测试集。通过记录 SI+DGC 处理过程在每个测试子集的识别正确率和召回率，得到了图 1 的结果。



(a) SI+DGC 的测试精度比较　(b) SI+DGC 的召回率比较

图 1 SI+DGC 的正确率和召回率在开放和封闭环境下的结果对比

从图 1 可以发现：

(1)在相同的测试环境下，正确率与召回率的性能曲线具有相似性：在开放测试中，正确率和召回率都是在最初经历

（下转第 182 页）