HMM 模型和句法分析相结合的事件属性信息抽取

吴家皋[1,2],周凡坤[1,2],张雪英[3]

(1. 南京邮电大学计算机学院,江苏 南京 210003)
(2. 江苏省无线传感网高技术研究重点实验室,江苏 南京 210003)
(3. 南京师范大学虚拟地理环境教育部重点实验室,江苏 南京 210023)

[摘要] 自然语言处理技术是计算机科学领域与人工智能领域中的一个重要方向,其中信息抽取是近年来新兴起的一个研究领域. 由于汉语自身结构松散、语法语义灵活等特点,使得中文文本中信息抽取具有较大的难度. 本文提出句法分析和隐马尔科夫模型相结合的事件属性抽取方法,其主要思想是先利用句法分析对中文文本进行分析,将得到的句法结构交给隐马尔科夫模型进行学习得到一个抽取模型,然后再由此模型对中文文本进行抽取. 实验表明,该方法具有较高的准确率和召回率.

[关键词] 自然语言处理,中文文本信息抽取,隐马尔科夫模型,句法分析,触发词
[中图分类号]TP181 [文献标志码]A [文章编号]1001-4616(2014)01-0030-05

Research of the Extraction Method of Event Properties Based on the Combining of HMM and Syntactic Analysis

Wu[1] Jiagao[1,2], Zhou Fankun[1,2], Zhang Xueying[3]

(1. School of Computer Science & Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)
(2. Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing 210003 China)
(3. MOE Key Laboratory of Virtual Geographic Environment, Nanjing Normal University, Nanjing 210023, China)

Abstract: Natural language processing technology is an important direction in the field of computer science and artificial intelligence, and the Chinese text information extraction is a new rising researching field in recent years. Due to the character of the loose structure of Chinese text, the flexibility of grammar and semanteme, the research of Chinese natural language processing has a difficult challenge nowadays. In the paper, a method of the combine of syntactic and HMM (Hidden Markov Model) was proposed. The main idea is to use syntax to analyze the Chinese text, then submit the syntactic structure to HMM and get a HMM model through learning it, finally the event properties can be extracted by the HMM model. The experiment shows that the method has higher precision and recall than normal algorithm.

Key words: natural language processing, information extraction of Chinese text, hidden markov model, syntactic analysis, trigger words

随着计算机在各个领域的广泛应用以及 Internet 的迅猛发展,越来越多的事件信息是以电子文档的形式在计算机中存储和处理. 为了能够及时有效地从中文文本中提取事件的属性信息,迫切需要一些自动化技术帮助人们在海量信息中找到自己真正需要的信息. 因此自动事件属性信息抽取是一个非常重要的研究课题.

目前信息抽取模型主要可以分为两大类:基于规则的抽取模型、基于统计的抽取模型. 隐马尔科夫模型(HMM)的基本理论是 20 世纪 60 年代末、70 年代初由 Baum 等人创建的. 它是一种机器学习的方法,它已经被广泛应用到各个领域,其中由于其简单、易建立、适应性强等特点,在事件属性信息抽取方面已得到广泛应用. Jiang Huixing 提出了采用一个新型的整形二阶 HMM 模型来抽取事件实体,但还是存在文本本身特征的不足引起抽取精度不足的问题[1]. Zhou Deyu 提出了使用 HVS(Hidden Vector State Model)模型

收稿日期:2013-08-10.
基金项目:国家 863 项目(2012AA12A403)、江苏省自然科学基金(BK2012833)、江苏省高校自然科学基金(12KJB520011).
通讯联系人:吴家皋,博士,副教授,研究方向:计算机网络、移动计算、GIS 应用等. E-mail:jgwu@njupt.edu.cn

来抽取生物学方面的事件属性信息,但是利用此模型的抽取效果仍然很局限[2]. 文献[3]提出了使用统计学的方法对文章中出现的主观情感片段进行抽取的方法. 文献[4]利用主动学习技术来减少训练 HMM 信息抽取模型时所需的标记数据,但是数学模型还不够完善,易陷入局部最优解,精确度还有待提高. 文献[5]提出了一种平滑发散概率函数和后向收缩技术以改进 HMM 模型,使得 HMM 模型提取元数据模型更加准确,但是抽取模型在提取的内容上有很大局限性.

句法分析是自然语言处理研究的关键性问题之一,其主要任务是自动识别句子的句法结构,即句子包含的句法单位以及这些句法单位相互之间的关系. 句法分析能够应用在很多方面,文献[6]给出了句法分析在中文分词领域中的应用. 句法分析是从单词串得到句子结构的过程,句法分析的最终目标是对于给定的句子,生成一个带有句法功能标记的短语结构树,句法分析的过程也可以理解为句法树的构造过程. 斯坦福大学和中科院均对中文语句的句法分析有一定的研究和成果. 随着自然语言应用的日益广泛,特别是对文本的处理需求的进一步增加,句法分析的作用将会愈加突出[7].

单独地使用 HMM 和句法分析模型也很难得到一个很好的效果. 文献[8]提出了采用句法分析模型对汉语定语语句的抽取,但是该方法抽取的条件和精确度有限. 文献[9]提出了对传统 HMM 模型的改进方法,考虑了对观测值的后向依赖,但是对预测结果精度的提高有限. 在本文中,为了提高抽取精度,我们引入词典,词典中包含有信息抽取所需的触发词,由于触发词和事件属性信息有着很高的相关度,所以能极大地提升抽取效果. 实验表明,该算法具有很好的精确度和召回率.

1 HMM 模型和触发词

1.1 HMM 模型

HMM 提供了一种基于训练数自动构造识别系统的技术. 一个 HMM 包含两层:一个可观察层和一个隐藏层. 可观察层是待识别的观察符号序列,隐藏层是一个马尔科夫过程. 一个一阶 HMM 模型 θ 可以看成一个五元组 $\theta=(N,M,\boldsymbol{A},\boldsymbol{B},\boldsymbol{p})$,每个字母解释如下:

(1)N:状态数目,设状态集合为 $S=\{s_1,s_2,\cdots,s_N\}$;

(2)M:观察符号数目,设观察符号集合为 $E=\{e_1,e_2,\cdots,e_M\}$;

(3)$\boldsymbol{A}=\begin{bmatrix} a_{s_1,s_1} & a_{s_1,s_2} & \cdots & a_{s_1,s_N} \\ a_{s_2,s_1} & a_{s_2,s_2} & \cdots & a_{s_2,s_N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{s_N,s_1} & a_{s_N,s_2} & \cdots & a_{s_N,s_N} \end{bmatrix}$:状态转移矩阵,其中 a_{s_i,s_j}(简记为 $a_{i,j}$)表示从状态 s_i 转移到状态 s_j 的概率;

(4)$\boldsymbol{B}=\begin{bmatrix} b_{s_1,e_1} & b_{s_1,e_2} & \cdots & b_{s_1,e_M} \\ b_{s_2,e_1} & b_{s_2,e_2} & \cdots & b_{s_2,e_M} \\ \vdots & \vdots & \ddots & \vdots \\ b_{s_N,e_1} & b_{s_N,e_2} & \cdots & b_{s_N,e_M} \end{bmatrix}$:观察符号概率矩阵,其中 b_{s_i,e_j}(简记为 $b_{i,j}$)表示当处于状态 s_i 时观察到符号 e_j 的概率;

(5)$\boldsymbol{p}=(p_{s_1},p_{s_2},\cdots,p_{s_N})$:初始状态概率矢量,其中 p_{s_i}(简记为 p_i)表示初始选取的状态为 s_i 的概率.

应用 HMM 模型,主要解决以下 3 个方面的问题:评估问题、学习问题和解码问题. 事件信息抽取需要解决 HMM 模型中的学习问题和解码问题[10]. 通过一定的算法得到这 5 个参数的值,HMM 模型的学习问题就解决了,然后利用此模型对观察符号序列进行解码,就可以得到状态序列,即可以识别出要抽取的事件属性信息.

1.2 触发词

触发词是相对于特定的主题事件,描述该事件的特定属性信息非常重要的结构组成. 例如在描述地震事件的中文文本中,其中“此次地震震级为 7 级”这句话中根据“震级”一词就能很容易地判断出“震级为 7 级”是我们在此文本中对于地震事件要抽取的属性信息. 虽然并不是所有特定主题事件中要抽取的属性信息中都含有触发词,但是含有触发词则更有可能成为要抽取的对象. 将触发词进行汇总分类然后整理成

触发词词典,然后我们就可以将此词典作为知识库,为事件属性信息的抽取提供先验知识.

2　HMM 模型与句法分析相结合的事件属性信息抽取

2.1　算法流程

对事件属性信息进行抽取时,可以建立很多种 HMM 模型,现在一般的模型是以单词或者词性作为可观测层的基本观测单元. 如若将单词作为观测层的观测单元,由于单词量的过于庞大,则会造成 HMM 模型观测符号过多,维数过大,导致模型不易建立并且模型预测结果失真. 采用词性作为观测层的基本观察单元,HMM 模型的观测符号不会由于过多而造成上述情况,经试验验证确实能得到不错的效果.

由于句法分析能提供更多的句子信息,所以将产生的句子结构标记作为 HMM 模型的观测序列,同时加入触发词,将会得到更好的识别效果. 本文采用斯坦福大学的句法分析软件进行实现. 句法分析与 HMM 模型相结合的抽取系统处理过程如图 1 所示.

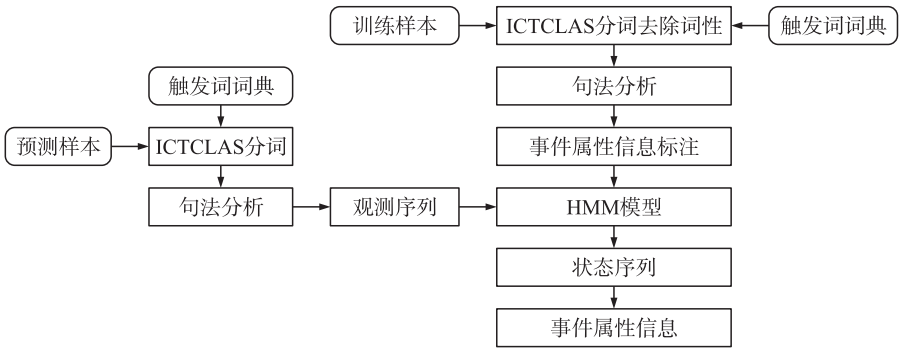

图 1　句法分析与 HMM 模型相结合的抽取系统处理过程

Fig. 1　Process of the extraction system based on the combinational model of HMM and syntactic analysis

由图 1 可以看出,整个流程分为两个步骤:学习过程和预测过程. 对训练样本学习时,首先利用 ICTCLAS 模型对训练样本进行分词,然后对分词结果进行句法分析和触发词标注,将句法分析得到的句法结构进行手工标注出要抽取的事件属性信息即状态标注,最后对标注后的文本利用 HMM 模型学习,得到一个 HMM 模板,这样就可以利用此模板对预测样本进行预测了. 对预测样本预测的预处理的过程与学习过程相同,同样要将文本利用句法分析模型和触发词词典得到中文文本的句法结构,然后利用 Viterbi 算法计算,被标记为目标状态标签的对应观察文本即为要抽取的事件属性信息.

2.2　HMM 模型的参数学习

对于 HMM 模型的建立一般采用 ML 算法(最大似然算法)(对于已标记的训练文本)和 Baum-Welch 算法(对于未标记的训练样本)进行学习. 此处训练文本已被标记,所以我们采用 ML 算法获取 HMM 模型.

Step 1　文本预处理:对文本进行句法标记.

将训练文本交给斯坦福大学句法分析器,对得到的文本进行标记,标记出要抽取的文本块. 同时借助触发词词典将文本中出现的触发词用特殊的标记标注出.

Step 2　应用 ML 算法建立 HMM 模型.

(1)N 的获取. 同上,文本信息分为被抽取的事件属性信息部分和不被抽取的其他部分,此处 $N=2$.

(2)M 观察符号数目的获取. 首先采用 ICTCLAS 分词工具对中文文本进行分词,此处不需要词性的标注,然后将分词交给斯坦福中文句法分析模型获取句子结构. 经统计,斯坦福大学的中文句法分析模型中总共有 48 个符号,代表不同的句子结构. 同上引入字典,加入触发词,当一个词是触发词时,这个词之前的符号就被标注为触发词类型,例如未加入字典前“死亡 3 人”句法分析后为“VP VV 死亡 NP QP CD 3 NP NN 人”,加入字典后为“VP VV_new 死亡 NP QP CD 3 NP NN 人”,VV_new 就是表明该词为触发词类型,由于每个符号都有可能被标注为触发词类型,所以此时共有 96 个观察符号,$M=96$.

(3)状态转移矩阵 $\boldsymbol{A}$、观察符号概率矩阵 $\boldsymbol{B}$ 与初始状态概率 $\boldsymbol{p}$ 的获取.

ML 算法以统计的方法获取 HMM 模型参数,则状态转移矩阵中每个参数通过下公式得到,

$$a_{ij}=\frac{C_{i,j}}{\sum_{k=1}^{N}C_{i,k}},1\leqslant i,j\leqslant N, \tag{1}$$

其中 $C_{i,j}$是对训练样本进行状态标注时,计算所有符号对应的状态从状态 S_i 转换到状态 S_j 的次数.

$$b_j(V_k)=\frac{E_j(V_k)}{\sum_{i=1}^{M}E_j(V_i)},1\leqslant j\leqslant N,1\leqslant i\leqslant M, \tag{2}$$

其中 $E_j(V_k)$是观察符号 V_k 在状态 S_i 中的发射次数,则 $b_j(V_k)$即是观测符号 V_k 在状态 S_i 中所有观测符号发射概率之和的比重.

$$p_i=\frac{\mathrm{Init}(i)}{\sum_{j=1}^{N}\mathrm{Init}(j)},1\leqslant i\leqslant N, \tag{3}$$

其中 Init(i)是以状态 i 开始的总个数.

2.3 基于句法分析与 HMM 模型相结合的属性抽取

Step 1 文本预处理.

将文本作为斯坦福大学句法分析程序进行句法标注,同时利用触发词字典标注出触发词汇.

Step 2 利用 Viterbi 算法得到状态序列.

应用已建立好的 HMM 模型对得到的标注序列 $O=O_1O_2\cdots O_n$ 采用 Viterbi 算法,找出最大概率的状态标签序列,此时被标记为目标状态标签的对应观察文本即为要抽取的内容.

3 实验

3.1 实验数据集和实验环境

本实验所使用的是从网络上摘取的 200 篇有关地震事件的文本预料. 将其中 180 篇作为训练文本,将其余的 20 篇作为测试文本.

实验 PC 机为:Acer Inter P6200,主频为 2.2 GHz,java 环境为 jdk1.6,采用 eclipse 作为开发平台.

3.2 实验评价准则

为评价抽取效果,我们采用最通用的性能评价方法:召回率 R(Recall)、准确率 P(Precision)、F_1 测度. 定义如下:

召回率:$R=\dfrac{\text{正确抽取的属性个数}}{\text{文本中所应抽取的属性个数之和}}\times 100\%$.

准确率:$P=\dfrac{\text{正确抽取的属性个数}}{\text{实际抽取到的属性个数}}\times 100\%$.

定义二维混淆矩阵,如表 1 所示.

$$R=\frac{A}{A+C}\times 100\%,P=\frac{A}{A+B}\times 100\%,$$

则 F_1 测度用如下公式获得:

$$F_1=\frac{2\times R\times P}{R+P}\times 100\%.$$

表 1 性能指标计算

Table 1 Performance metrics calculation

	应该抽取的属性个数	不应该抽取的属性个数
实际抽取得到的属性个数	A	B
未抽取的属性个数	C	D

3.3 实验结果分析

分别使用利用词性标注的 HMM 抽取模型,词性标注和触发词词典相结合的 HMM 模型与句法分析与 HMM 模型相结合的抽取模型,同时与文献[10]中的利用文本中能反应句子结构的特征词作为观测符号的模型对相同的文本集合做测试比较. 从表 2 可以看出,HMM 模型与句法分析相结合的模型相比较而言具有更高的准确率和召回率,并且触发词能极大地提高 F_1 值,说明通过句法分析和触发词可以得到更多句子结构信息,对于文本属性信息的抽取起着不可忽视的作用.

由图 2 可以看出,随着训练样本的增加,中文文本属性抽取的效果也会随之增加,但是也会随之趋于稳定. 这是由于当一个 HMM 模型建立好以后,因为中文文本语言描述千差万别,一个固定的模型只能保

证此模型能照顾到大部分常见的句子结构,但毕竟中文文本语言描述千差万别,一个模型很难将所有的情况概括进来. 同时从图 2 可以看出,HMM 模型与句法分析相结合的算法比上述文中提到的 HMM 模型的抽取效果有了一定的改善.

表 2　实验结果

Table 2　Experimental results

	准确率 P/%	召回率 R/%	F_1 值/%
仅词性标注	56.329	54.268	55.279
反映句子结构的特征词	62.312	64.266	63.313
词性标注和触发词词典结合	74.624	91.463	82.190
句法分析与 HMM 模型相结合	80.864	92.254	86.184

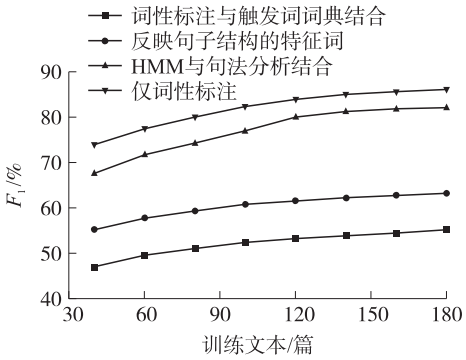

图 2　传统模型与 HMM 和句法分析模型 F_1 值的对比

Fig. 2　Comparisons of F_1 with traditional models and the combinational model of HMM and syntactic analysis

4　结语

针对传统的基于 HMM 模型的事件属性抽取方法的不足,本文提出了基于句法分析与 HMM 模型相结合的事件属性信息抽取方法,同时在这种方法中引入了触发词词典. 实验表明,加入触发词之后,能够极大地提高事件属性信息抽取的准确率和召回率,并且在考虑了句子结构信息之后,又进一步地提高了准确率和召回率,在一定程度上克服了传统方法的不足,但该模型仍有不完善之处,下一步将对算法作进一步的优化,可适当地考虑多模板以及增加 HMM 模型的阶数,进一步提高模型的准确率及其适用性.

[参考文献]

[1] Jiang Huixing, Wang Xiaojie, Tian Jilei. Second-order HMM for event extraction from short message[J]. Lecture Notes in Computer Science, 2010, 6 177: 149-156.

[2] Zhou Deyu, Yulan Heb. Biomedical events extraction using the hidden vector state model[J]. Artificial Intelligence in Medicine, 2011, 53(3): 205-213.

[3] Li Qing, Yuanzhu Peter Chen. Personalized text snippet extraction using statistical language models[J]. Pattern Recognition, 2010, 43(1): 378-386.

[4] Scheffer T, Decomain C, Wrobel S. Active hidden Markov models for information extraction[C]//Proceedings of the International Symposium on Intelligent Data Analysis. Berlin: Springer, 2001: 301-109.

[5] Bolanle Ojokoh, Zhang Ming, Tang Jian. A trigram hidden Markov model for metadata extraction from heterogeneous references[J]. Information Sciences, 2011, 181(9): 1 538-1 551.

[6] Liu Jiangyang. Resolution to combinational ambiguity of Chinese word segmentation[C]//Proceedings of the International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government, 2009. Hongkong: IEEE, 2009: 141-145.

[7] 袁里驰. 基于依存关系的句法分析统计模型[J]. 中南大学学报:自然科学版, 2009, 40(6): 1 630-1 635.

[8] 张宴生. 汉语定义语句的抽取方法[J]. 计算机与数字工程, 2011, 39(10): 45-47.

[9] 梁吉光, 田俊华, 姜杰. 基于改进 HMM 的文本信息抽取模型[J]. 计算机工程, 2011, 37(20): 178-182.

[10] Souyma Ray, Mark Craven. Representing sentence structure in hidden markov models for information extraction[C]// Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence. Washington: Morgan Kaufmann Publishers, 2001: 1 273-1 279.

[责任编辑:丁　蓉]