doi: 10.3969/j.issn.1007-855x.2009.04.007

基于决策树的现代汉语中任职关系抽取研究

帅训波[1],马书南[2]
(1.中国石油勘探开发研究院 廊坊分院 地球物理与信息研究所,河北 廊坊 065007;
2.北京工业大学 计算机学院,北京 100022)

摘要:在命名实体识别的研究基础之上,论文把抽取人名实体与机构实体间的任职关系看成分类问题.即根据现代汉语句子中任职动词的类别属性将任职关系信息抽取模式分类.应用决策树的方法确定句子的抽取模式,实现人在机构中的任职关系信息抽取.并对建立的基于该决策树的任职关系抽取系统进行开放测试,平均召回率和精确率分别为 91.47%和 89.15%,实验结果表明,基于决策树的现代汉语中任职关系抽取是一种值得继续探讨的方法.

关键词:命名实体识别;决策树;信息抽取;自然语言处理
中图分类号:TP391 **文献标识码**:A **文章编号**:1007-855X(2009)04-0027-05

Research of Duty Information Extraction in Chinese Based on Decision Tree

SHUAI Xun-bo[1], MA Shu-nan[2]
(1. Institute of Geophysics and Information, Langfang Branch of Research Institute of Petroleum Exploration and Development, Langfang, Hebei 065007, China;
2. College of Computer Science, Beijing University of Technology, Beijing 100022, China)

Abstract: Based on the named entity recognition research, duty information extraction is taken as a question of classification first, that is, information extraction mode is classified by duty verb in Chinese sentences. Decision tree is used to select appropriate extraction mode, which solves the problem of duty information extraction. A Chinese duty information extraction system based on the decision tree is thus realized. For an open test, its average recall rate is 91.47% and average precision is 89.15%. Experimental results show that the information extraction method based on decision tree is worth to continue for further research.

Key words: named entity recognition; decision tree; information extraction; natural language processing

0 引 言

关系信息抽取是指从文档中识别出实体之间或实体及其属性之间的关系.随着命名实体识别和职务实体识别的研究逐步深入[1~5],对现代汉语中实体间关系抽取的研究也越来越得到重视[6,9].人与机构间任职关系是现代汉语中较常见的信息,实现现代汉语中人与机构间的任职关系信息抽取具有重要的意义[8,9].基于统计或机器学习算法的角度,进行了现代汉语中实体关系抽取研究,由于现代汉语句子自身在结构和语义等方面的复杂度差异,构造全面而有效的抽取算法或模型是难以实现的.本文在命名实体识别研究基础之上,从语言学的角度,将人名实体与机构实体间的任职关系看成分类问题.即根据现代汉语句子中任职动词的类别属性,将任职关系信息抽取模式分类.应用决策树的方法确定句子的抽取模式,实现人在机构中的任职信息抽取.并且对建立的基于该决策树的任职关系抽取系统进行开放测试,平均召回

收稿日期:2008-11-03.**基金项目**:河北省科学技术进步成果资助项目(项目编号:20070305).
第一作者简介:帅训波(1979-),男,硕士,工程师.主要研究方向:进化计算、人工智能、数据挖掘等.**E-mail:** sxblffy@126.com

率为91.47%,平均精确率为89.15%,实验结果表明,基于决策树的现代汉语中任职关系抽取是一种值得继续探讨的方法.

1 决策树的构建

1.1 问题的描述

命名实体识别方法是研究的基础,也就是说,在已经标注了人名和机构名的现代汉语文本中进行人与机构间任职关系的信息抽取,将核心抽取过程分成2个阶段:①对职务词和任职动词的正确识别;②基于决策树的信息抽取模式应用;现代汉语中人与机构间的任职关系有当前任职、职务建立、免职和职务变化等多种状态,它们主要通过句子中的不同任职动词体现出来,这也是我们对任职动词属性进行详细分类的主要依据.因此,我们在完成文本中职务词和任职动词的识别之后,再进一步根据任职动词的类别属性查询决策树,选取正确的抽取模式,实现人在机构间的任职关系信息抽取.

1.2 职务实体和任职动词识别

首先建立职务核心词表和任职动词表,对现代汉语中职务核心词和任职动词进行收录.通过对职务核心词表和任职动词表的快速检索,实现对现代汉语中职务实体和任职动词的识别.

在现代汉语中有相当数量的职务表述由核心职务词与其前缀词或后缀词组成,例如:"副所长"、"总经理助理"等诸类职务词由核心职务词"所长"、"总经理"分别和前缀词"副"、后缀词"助理"组合表述构成.核心词数据项与前缀词数据项或后缀词数据项组合对现代汉语句子中职务实体准确识别.对句子中的职务词识别检索时,先检索匹配职务核心词,然后再对职务核心词的前缀词或后缀词进行检索,进一步准确识别职务词.为了加快对句子中职务词的检索识别速度,在职务核心词表中增加被检索概率数据项,依据对职务核心词被检索成功的概率的大小,对职务核心词表进行排序,有效利用前文检索的"遗留"信息,为下次对职务核心词表的检索提供"优越"条件,从而使得在历史上被检索成功率较高的职务核心词以较快的速度被检索.职务核心词表结构如表1所示:

表1 职务核心词表

Tab.1 Key words of duty entity

被检索概率	核心词	前缀词	后缀词
70%	总经理	副、常务副、原	助理、秘书
58%	主任	副、正	
—	—	—	—

依据现代汉语中任职动词的语义,可以分为职务建立类(例如:出任、任命等)、职务解除类(例如:免去、罢免等)和职务变更类(例如:调离、调动等).任职动词在现代汉语句子中对句子的构成起到重要作用,通常有基本句式表述,也可以以基本句式为基础进行变化,换成另外一种表述句式.例如"[李小波][出任][院办][副主任]"的表述为任职动词"出任"的基本式,又可以变化为"[院办][副主任]由[李小波][出任]"的方式.我们借鉴职务核心词识别时构建词表的方法,通过检索任职动词表,实现对任职动词的识别.增加任职动词在句子中可能出现的模式项,利用已经识别的人名、机构名和职务词的信息,来进一步确认检索的任职动词是否为抽取任职关系的动词.构建的任职动词表如表2所示:

表2 任职动词表

Tab.2 Verbs of duty appointment

被检索概率	动词	性质	基本式	变化式
88%	担任	建立	[人名]担任[机构]的[职务]	[机构]的[职务]由[人名]担任
63%	免去	解除	免去[人名][机构][职务]	[人名][机构][职务]被免去
—	—	—	—	—

1.3 任职信息抽取模式

在现代汉语中,对人物职务身份表达时,除了应用任职动词明确陈述之外,不含任职动词的静态任职

关系表述也是另外一种常见的句式.这种表述有两种基本形式:“是”字句和同位关系短语.“是”字句作为一种特殊句式,例如:“年仅 32岁的李晓明是北京神州九天公司的总经理”等类句子.这类句子以“是”为核心,均为“[人名]是[机构名][职务词]”的模式.作为句子成分短语是对任职关系静态表达的另一种较普遍情况,这种短语的通常以“[机构名][职务词][人名]”的习惯模式出现在句子中,例如:“山东省单县三中校长隋永迁出席第十七次全国人民代表大会”.我们将现代汉语句子依据是否含有任职动词进行分类,然后再对不含任职动词的句子根据各自句式结构特点进行细分类;对含任职动词的句子,根据任职动词及其表述模式进行详细分类;经过细分后的每一类均是对任职关系进行明确抽取的模式,根据抽取模式进行任职关系信息抽取.

1.4 决策树建立

决策树把客观世界或对象抽象为一个信息系统,也称属性-值系统.这个信息系统 S是一个四元组:$S = < U, A, V, f >$,其中,U是一组对象(或事例)的有限集合,称论域;设有 n个对象,则 U可表示为:$U = \{x_1, x_2, x_3, \cdots, x_n\}$;$A$是有限个属性的有限集合,设有 m个属性,则其可表示为:$A = \{a_1, a_2, \cdots, a_m\}$;$V$是属性的值域集,$U = \{V_1, V_2, V_3, \cdots, V_m\}$,其中 V_i是属性 a_i的值域;A又可进一步划分为 2个不相交的集合:描述属性集 C和决策属性集 D,C和 D满足 $A = C \cup D$且 $C \cap D = \phi$,D一般只有 1个属性.f是信息函数,$f: U \times A \rightarrow V$, $f(x_i, a_j) \in V_j$.

基于这种四元组构造的决策树分类器,输入是一组带有类别标记的例子,即对象集 U、对象的描述属性集 C、决策属性 D和与之相对应的值域集 V,构造的结果是一棵二叉树或多叉树.树的内部节点(非叶子节点)一般表示为一个逻辑判断;树的边是逻辑判断的分支结果,对于离散属性,内部节点是属性,边是该属性的所有取值,有几个属性值,就有几条边.树的叶子节点都是类别标记,即决策属性 D的取值.

将信息抽取的现代汉语句子作为决策树的输入对象,描述属性集 $C = \{$人名、机构名、职务词、任职动词$\}$,任职动词的性质和在句中的表述模式等为构造决策属性值,决定任职关系信息抽取的模式,将 C中各元素遵循抽取模式进行任职关系信息抽取.决策树的构造结果如图 1所示:

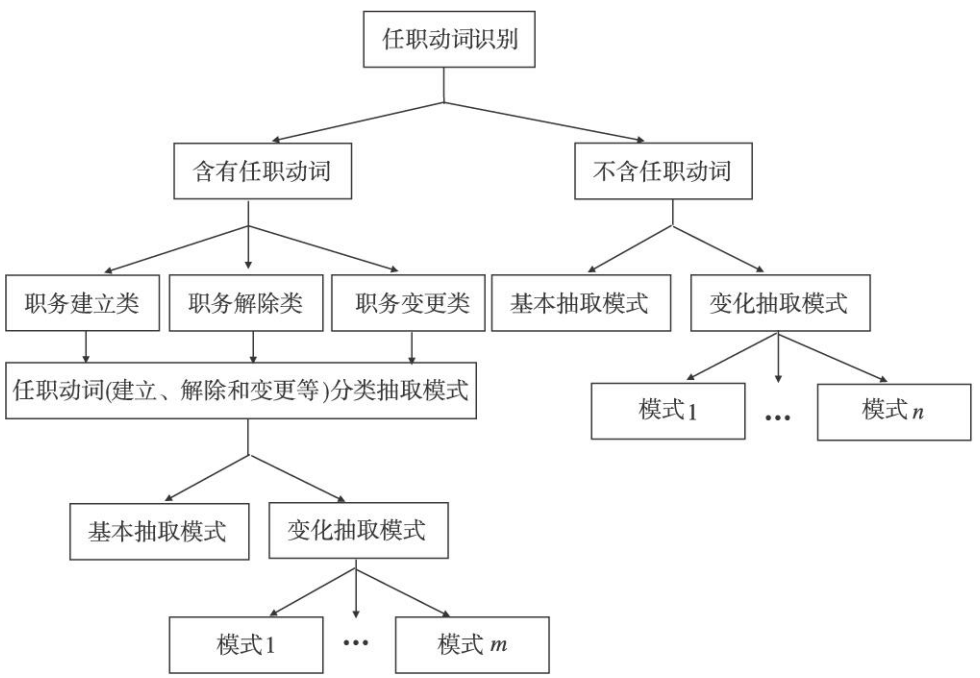

图 1　决策树构建

Fig. 1　Construction of decision tree

2 基于决策树的任职信息抽取系统

2.1 系统模型

借鉴 Hobbs提出的信息抽取系统体系结构思想[10],设计基于上述决策树的任职关系信息抽取系统结构模型如图 2所示:

1) 文本分块:将输入的现代汉语文本分割为不同的块,系统主要是以自然段落的形式对文本进行分割;

2) 文本预处理:将得到的文本块转换为句子序列,对每个句子进行人名实体和机构实体识别[1-6],对句中的人名和机构名进行标注;

3) 文本过滤:对于文本分块中,提取需要抽取任职关系的句子,如果句子缺少人名实体而有人称代词,或缺少机构实体而有机构指示代词的情况下,将代词还原为相应的人名实体或机构实体,对句子进行规范化;过滤掉不相关的句子;

4) 任职关系抽取:对职务核心词表进行检索,查找句子中的所有职务词;对任职动词表进行检索,识别句子所有任职动词,应用本文构建的抽取决策树对句子进行任职关系实例抽取;

5)抽取结果:根据抽取实例,得出人与机构间的任职关系,并且动态更新任职动词表和职务核心词表中的被检索概率项.

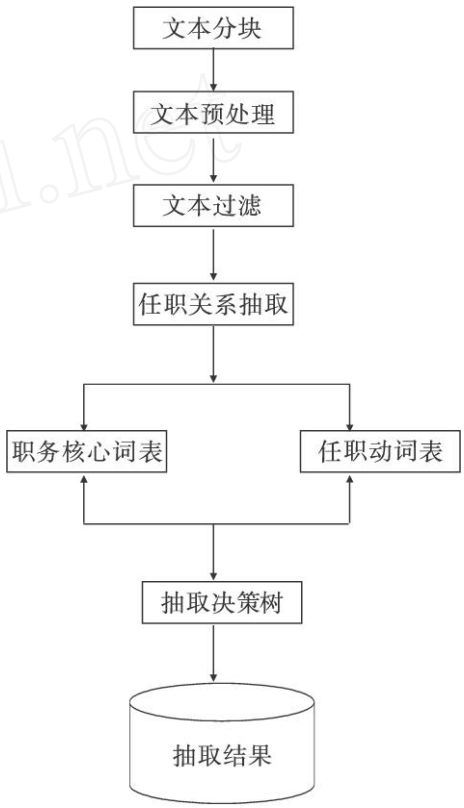

图 2 基于决策树的任职关系信息抽取系统模型

Fig. 2 Model of duty information extraction system based on decision tree

2.2 抽取过程

本文基于决策树的任职关系抽取过程描述如下:

1) 对需要处理的文本进行分块,获取抽取处理的文本句子;

2) 对句子进行预处理和过虑,抽取句子人名实体和职务实体;

3) 通过检索职务核心词表和任职动词表,对句子中的职务词和任职动词进行识别;

4) 依据任职动词分类,查找抽取决策树,获取信息抽取模式;

5) 将人名、机构名、职务词和任职动词根据信息抽取模式,进行实例化,得到人在机构中任职关系的信息抽取结果;

6) 更新任职动词表和职务核心词表;

7) 判断是否还有需要抽取处理的文本句子,如果全部文本句子处理完毕,则对抽取的任职关系进行信息融合;否则主则转向步骤 2).

2.3 系统实现的关键技术

职务核心词表和任职动词表以二维表格的形式存在数据库中. 任职动词表和职务核心词表主要是由人工经验归纳完成,任职动词表和职务核心词表领域范围和词的量均需要具有一定的规模,尽可能地覆盖常用职务核心词和任职动词,以保障任职关系信息抽取系统有较好的抽取准确率.

抽取决策树通过树形数据结构实现,对每一条从根结点到叶子结点的遍历路径均可得到一种抽取模式. 对每类任职动词在现代汉语中习惯用到的表述方式尽可能地包括在决策树叶子结点集合中. 在过滤过程中的人称或机构指示代词分别还原成人名实体或机构实体时,按照上下文本句子间的就近原则进行处理,这也是符合现代汉语表述习惯的.

3 实验结果

为了检验系统的任职关系抽取性能,应用 Java语言编程实现,随机抽取 20篇新华社新闻稿件作为每次实验的实验数据,在 Windows环境下对其测试,得到令人满意的效果. 采用对系统性能进行评价[6]. 其定义如下:

$$F\text{-}Score = \frac{2 * Precision * Recall}{Precesion + Recall};$$

其中准确率 (Precision) 和召回率 (Recall) 的定义为:

$$Precision = \frac{\text{被正确抽取的关系}}{\text{抽取的所有关系}} \times 100\%;$$

$$Recall = \frac{\text{被正确抽取的关系}}{\text{测试数据的所有任职关系}} \times 100;$$

随机抽取其中 10次实验结果,求 P*recision*和 R*ecal*的平均值分别为 89.15%和 91.47%. 实验结果表明,本文所提出的基于决策树的任职关系信息抽取方法是可行的,召回率的提高与任职动词表和职务核心词表的规范建设有密切关系.

4 结 论

从任职动词在现代汉语句子结构中的重要作用入手,把抽取人名实体与机构实体间的任职关系看成分类问题,根据任职动词类别属性确定人在机构间的任职关系抽取表述模式,建立了模式抽取决策树,并对基于该决策树的任职关系抽取系统进行实验. 开放测试的实验结果表明,论文所提出的基于现代汉语句子中任职动词的语言属性进行人在机构间任职关系抽取方法是有效的. 同时,也为应用现代汉语句子自身表述特点与计算机智能识别相结合,实现对现代汉语中的信息抽取研究提供一定的借鉴意义.

参考文献:

[1] 沈嘉懿,李芳,徐飞玉,等. 中文组织机构名称与简称的识别 [J]. 中文信息学报, 2007, 21(6): 17 - 21.
[2] 冯冲,陈肇雄,黄河燕. 采用主动学习策略的组织机构名识别 [J]. 小型微型计算机系统, 2006, 27(4): 710 - 714.
[3] 张晓艳,王挺,陈火旺. 命名实体识别研究 [J]. 计算机科学, 2005, 32(4): 44 - 48.
[4] 王振华,孔祥龙,陆汝占,等. 结合决策树方法的中文姓名识别 [J]. 中文信息学报, 2004, 18(6): 10 - 15.
[5] 张华平,刘群. 基于角色标注的中国人名自动识别研究 [J]. 计算机学报, 2004, 27(1): 85 - 91.
[6] 车万翔,刘挺,李生. 实体关系自动抽取 [J]. 中文信息学报, 2005, 19(2): 1 - 6.
[7] 熊邓攀,樊孝忠,杨立公. 用语义模式提取实体关系的方法 [J]. 计算机工程, 2007, 33(10): 212 - 214.
[8] 刘克彬,李芳,刘磊,等. 基于核函数中文关系自动抽取系统的实现 [J]. 计算机应用研究, 2007, 44(8): 1406 - 1411.
[9] 何婷婷,徐超,李静,等. 基于种子自扩展的命名实体关系抽取方法 [J]. 计算机工程, 2006, 32(1): 183 - 185.
[10] HOBBS J. The Generic Information Extracting System [C]. In Proceeding of the Fifth Message Understanding Conference, Morgan Kaufman, 1993: 87 - 91.