专题研究

面向项目申请书的命名实体抽取模型构建研究

王文龙[1,3] 王东波[2,3]

[[1]南京大学信息管理学院 江苏 210093;
[2]南京农业大学信息科学技术学院 江苏 210095;
[3]江苏省数据工程与知识服务重点实验室(南京大学) 南京 210093]

摘要 文章总结了项目申请书中命名实体的分布特点,并根据这种分布特点,利用条件随机场构建了面向项目申请书的命名实体抽取模型,并对模型的性能进行了验证。实验表明,模型能够较好地对项目申请书中的命名实体进行抽取。

关键词 申请书 命名实体 条件随机场

Project Application- oriented Named Entity Extraction Model Construction

Wang Wenlong[1,3] Wang Dongbo[2,3]

[[1]School of Information Management, Nanjing University, Jiangsu, 210093;
[2]School of Information Science Technology, Nanjing Agricultural University, Jiangsu, 210095;
[3]Jiangsu Key Laboratory of Data Engineering and Knowledge Service(Nanjing University), Nanjing, 210093]

Abstract This paper summarizes the distribution characteristic of named entities in project application. To extract name entities in project application, the authors construct a CRFs- based model and experiment on performance of this model, the experiment result shows that this model performs well on named entities extraction of project application.

Keywords project application, named entities, CRFs

1 引言

企业项目申请书是企业为了申请项目向项目主管部门所提交的企业申请意向和资质的说明。它是企业申报项目必填的材料,项目申请书填写的正确性和规范性往往对企业能否成功申报项目有很大的影响。而项目申请书填写的正确性往往取决于企业对项目申报通知的解读和理解程度。项目申报通知是企业进行项目申报的依据,它包含了非常重要的项目知识,如项目的名称、级别、对申报主体的要求等。但是由于项目申报通知往往是半结构化甚至非结构化的文本,从申报通知中发现并抽取这些项目知识往往需要人工手动进行,无法通过机器进行自动的解读。因此,对项目申报通知中的关键项目知识进行界定并抽取,有利于更好地利用机器对申报通知进行自动化解读,从而提高项目解读效率并更有针对性的帮助企业进行项目的申报,从而达到服务部门对企业进行知识服务的目的。本文将试图对项目申报通知(以下简称项目书)中的关键项目知识进行界定,并构建针对关键项目知识的命名实体抽取模型。

2 命名实体识别和抽取进展研究

命名实体的识别和抽取作为自然语言处理的基础性工作,受到了国内外学术界的广泛重视。MUC 和

本文系国家自然科学基金面上项目"面向知识服务的知识组织模式与应用研究"(编号:71273126)的研究成果之一。

ACE 等自然语言处理领域的测评会议吸引了研究者们开发信息抽取(IE)系统,并根据这些系统在举办方提供的测评文档库上的表现进行打分。在其中,命名实体的抽取都被作为一项独立的任务参与测评。国内"863 计划"命名实体评测小组在 2004 年度将命名实体识别作为一项独立的任务提出[1]。目前,对命名实体的抽取方法主要可以分为基于规则的方法和机器学习(基于统计)的方法。

基于规则的方法,主要是将词法、语法和领域背景下的语义等方面的规则进行总结,并在识别和抽取过程中加入这些规则,以期提高命名实体抽取的准确率。基于规则的方法优点是针对某个领域准确率高,代表性的有参与 MUC-6 测评的 Proteus 系统[2]、参与 MUC-7 测评的 Lasie-II[3]、NetOwl[4]系统等。国内也有学者针对中文命名实体的特点,提出了一些基于规则的方法,这些方法在人名识别、地名识别等方面都取得了一定的进展。例如,孙茂松等[5]利用字等分布信息和人名称谓信息等制定规则集对人名进行识别,取得了较高的精确率和召回率;吕雅娟[6]建立规则采用分解处理、规则综合比较等步骤,对人名、地名和国外译名进行识别。但是,由于基于规则的方法的抽取准确率在很大程度上取决于所指定规则的质量,所以,规则的编写往往需要具有领域知识和一定语言学背景的学者。基于规则的方法缺点是可移植性比较差,因为规则的制定往往要依赖具体的领域知识,所以当需要向不同的领域移植时,需要对规则做比较大的改动。

机器学习的方法,又叫做基于统计的方法,就是根据统计学的原理利用机器学习的方法对语料中的命名实体进行识别并抽取的方法,其中比较经典的方法有隐马尔科夫模型(Hidden Markov Model, HMM)、支持向量机(Support Vector Machine, SVM),最大熵(Maximum Entropy, ME)、条件随机场(Conditional Random Field, CRF)等。Bikel 等[7]提出基于 HMM 的英文命名实体识别方法,对 MUC-6 的测评数据进行识别,取得了较高的准确率和召回率。张华平等[8]将 HMM 模型应用到中文人名的识别当中,在《人民日报》语料库中进行应用,也有较高的准确率和召回率。文献[9-10]中将 SVM 方法用于命名实体识别,也都取得了不错的效果。McCallum 等[11]在 Conll-2003 会议上将 CRF 模型应用到命名实体识别的任务中,测评结果准确率和召回率比较高。

3 项目书中的命名实体分布情况

3.1 项目书中的关键项目知识集的确定

项目书中的关键项目知识是指项目申请书中对项目申请具有指导性或者约束性的项目知识,如项目的级别、申报要求等,能否满足这些条件往往决定了企业是否可以申报某个项目,能否成功申报项目。关键项目知识集是指项目申请书中所有关键项目知识的集合。一个完整的关键项目知识集必须包含企业申报项目所需要的全部信息,同时又要排除冗余信息,利用关键项目知识集能够更加准确的确定项目申报的所需条件,从而可以提高项目解读效率。

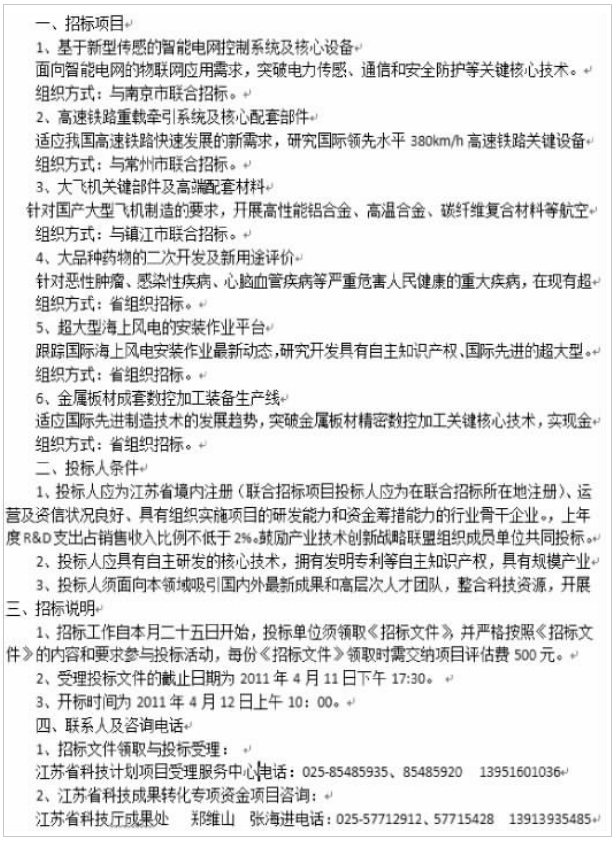
关键项目知识集的确定需要具有丰富项目申请经验的专家来确定。本文在研究项目申请书和咨询相关专家的基础上,确定了项目书中的关键项目知识集。它一共包含了 17 项关键项目知识,如表 1 所示。其中非粗体部分是项目书的基本信息,需要在项目书采集过程进行确定,而粗体部分为项目的必要知识,需要从项目书文件中进行抽取。由于从自然语言处理的角度,这些项目关键知识实际就是文本中的命名实体,所以,接下来本文将以命名实体的方法,把这些关键项目知识等价于命名实体,并构建模型重点进行抽取。

表 1 项目申请书中的关键项目知识集

3.2 项目书中的命名实体分布情况

项目书中命名实体的分布情况是指关键命名实体集中的命名实体在项目书中的位置分布情况。厘清命名实体的分布情况,可以帮助我们更有针对性地进行实体抽取。对命名实体的分布情况进行分析,是构建抽取模型的基础性工作。下页图 1 是本文将要进行命名实体抽取的某篇文献对象的部分,本文将以该文献为例,对待抽取项目书的文档结构和项目书中的命名实体的分布情况进行说明,以便后续的命名实体的抽取工作的进行。为了方便呈现,笔者在不影响命名实体抽取的前提下对它的篇幅进行了压缩。

从下页图 1 中可以看出命名实体在项目书中的分布与项目书的文档结构有着非常密切的联系。该篇公告的文档结构分为四大部分:招标项目、投标人条件、招标说明、联系人及咨询电话。针对该篇文档的命名实体的抽取方法为:在第一部分招标项目中,可以对项目名称、项目要求和组织方式等命名实体进行抽取;在第二部分投标人条件中,可以对项目的申报申报主体、申报主体要求、优先条件、限报条件等命名实



一、招标项目

1、基于新型传感的智能电网控制系统及核心设备

面向智能电网的物联网应用需求，突破电力传感、通信和安全防护等关键核心技术。

组织方式：与南京市联合招标。

2、高速铁路重载牵引系统及核心配套部件

适应我国高速铁路快速发展的新需求，研究国际领先水平 380km/h 高速铁路关键设备

组织方式：与常州市联合招标。

3、大飞机关键部件及高端配套材料

针对国产大型飞机制造的要求，开展高性能铝合金、高温合金、碳纤维复合材料等航空

组织方式：与镇江市联合招标。

4、大品种药物的二次开发及新用途评价

针对恶性肿瘤、感染性疾病、心脑血管疾病等严重危害人民健康的重大疾病，在现有超

组织方式：省组织招标。

5、超大型海上风电的安装作业平台

跟踪国际海上风电安装作业最新动态，研究开发具有自主知识产权、国际先进的超大型。

组织方式：省组织招标。

6、金属板材成套数控加工装备生产线

适应国际先进制造技术的发展趋势，突破金属板材精密数控加工关键核心技术，实现金

组织方式：省组织招标。

二、投标人条件

1、投标人应为江苏省境内注册（联合招标项目投标人应为在联合招标所在地注册）、运营及资信状况良好、具有组织实施项目的研发能力和资金筹措能力的行业骨干企业，上年度 R&D 支出占销售收入比例不低于 2%。鼓励产业技术创新战略联盟组织成员单位共同投标。

2、投标人应具有自主研发的核心技术，拥有发明专利等自主知识产权，具有规模产业

3、投标人须面向本领域吸引国内外最新成果和高层次人才团队，整合科技资源，开展

三、招标说明

1、招标工作自本月二十五日开始，投标单位须领取《招标文件》并严格按照《招标文件》的内容和要求参与投标活动，每份《招标文件》领取时需交纳项目评估费 500 元。

2、受理投标文件的截止日期为 2011 年 4 月 11 日下午 17:30。

3、开标时间为 2011 年 4 月 12 日上午 10：00。

四、联系人及咨询电话

1、招标文件领取与投标受理：

江苏省科技计划项目受理服务中心电话：025-85485935、85485920　13951601036

2、江苏省科技成果转化专项资金项目咨询：

江苏省科技厅成果处　郑维山　张海进电话：025-57712912、57715428　13913935485

图 1　科技项目书文档示例

体进行抽取；在第三部分招标说明中，可以对项目的申报方法、申报截止时间等命名实体进行抽取；在第四部分联系人及咨询电话中没有包含关键项目知识。所以，对项目的文档结构进行总结，可以方便更有针对性地对命名实体进行定位，对提高命名实体抽取效率有帮助。笔者通过对江苏省科技厅的近 5 年的项目书进行分析，发现项目书的文档结构与项目书的类型有很大的相关性，因此，首先要对项目书的文档类型进行总结，如表 2 所示。在此基础上，笔者针对每种文档类型总结了项目书中的命名实体分布情况，并对提取入口进行了初步归纳，以期对命名实体抽取模型的建立提供参考，如表 3 所示。其中，笔者对命名实体抽取入口相近的文档，采用了相同的命名实体抽取入口。

表 2　科技厅项目书文档类型总结

表 3　科技厅项目书文档结构及命名实体提取入口总结

4　项目申请书中命名实体的抽取

从本文第二部分可以看出，机器学习的方法由于其较好的可移植性目前在命名实体的识别和抽取中被普遍采用。而在连续文本的识别和抽取中比较常用的学习方法有最大熵模型（ME）、隐马尔科夫模型（HMM）和条件随机场模型（CRF）。其中，条件随机场模型不用像 HMM 模型那样需要非常严格的独立性假设，而且，条件随机场是从整体的角度进行决策，能够在前后序列元素之间做出平衡，而不像最大熵马尔科夫模型那样出现标记偏置，因此，条件随机场被很多学者认为是目前处理序列化数据标注的最好模型。故本文中将基于条件随机场（CRFs）构建抽取模型对申请书中的命名实体的识别。

4.1　条件随机场模型介绍

条件随机场模型（CRF）是一种判别图模型，由 Lafferty 等[13]于 2001 年提出。设 G=（V, E）是一个无向图，Y=｛Y_v| v ∈ V｝是以 G 中节点 v 为索引的随机变量 Y_v 构成的集合。在给定 X 的条件下，如果每个随机变量 Y_v 服从马尔科夫属性，则（X, Y）即构成一个条件随机场。令 X=｛x_1, x_2, …, x_i｝表示观察序列，Y=｛y_1, y_2, …, y_i｝表示观察序列对应的标记序列。

则如图 2 所示的条件随机场满足：

$$P(Y \mid X, \lambda) \propto \exp\left[\sum_j \lambda_j t_j\left(y_{i-1}, y_i, x, i\right) + \sum_j \mu_k s_k\left(y_i, x, i\right)\right] \quad (1)$$

其中 t（y_{i-1}, y_i, x, i），为对于观察序列的标记位置 i－1 与 i 之间的转移特征函数，S（y_i, x, i）为观察序列 ii 位置的状态特征函数，将两个特征函数统一为

$f(y_{i-1}, y_i, x, i)$，我们可以得到条件式：

$$P(Y \mid X, \lambda) = \frac{1}{Z(x)} \exp\left[\sum_{i=1}^{n}\sum_{j}\lambda_j f_j(y_{i-1}, y_i, x, i)\right] \quad (3)$$

其中，$Z(x) = \sum_{j}\exp\left[\sum_{i=1}^{n}\sum_{j}\lambda_j f_j(y_{i-1}, y_i, x, i)\right]$，公式（3）即为条件随机场的表示形式。

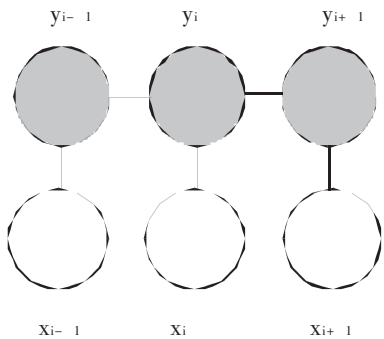


图 2 链式条件随机场

4.2 命名实体抽取中特征选择

CRFs 模型中，选择合适的特征对于取得良好的抽取效果非常关键。王春雨等[13]将命名实体的识别中使用的特征归纳为词汇特征，词法、句法特征和语义特征三类。其中，词汇特征为词语的外形的构成特点，如词中是否含有数字、字母等；词法、句法特征为词在句子结构中的成分和上下文信息，如词的词性等。对于比较复杂的命名实体，一般都会有特征词对某一类实体进行标识，例如公司、厅等词；还会有指示实体的左右边界词。本文在分析项目申请书中命名实体的外部特征和内部特征的基础上，最终选定的特征包括词本身、词性、边界词、命名实体特征词作为特征，如表 4 所示。

表 4 原子特征

同时本文在进行语料标注时采用了如下对命名实体的标注规则：B- PER、I- PER、E- PER、B- LOC、I- LOC、E- LOC、B- ORG、I- ORG、E- ORG、O 等。前面几个标注由横线相连的前后两部分组成，前一部分表示词语在命名实体中的位置：B 表示开始，I 表示内部，E 表示结束；后一部分表示命名实体的类别，PER 表示人名，LOC 表示地名，ORG 表示机构名。据此，B- PER 表示当前词是人名的首个词，I- PER 表示当前词是人名的中间词，E- PER 表示当前词是人名的结束词，其他类别表示规则相似。O 代表其他。部分标注语料格式如表 5 所示。

表 5 部分标注语料格式

4.3 命名实体抽取模型构建

（1）模型构建

命名实体抽取的流程主要是由训练和测试两部分组成。训练模块主要是在条件随机场的基础上，使用所确定的自身特征和添加特征模板，在训练语料上得到知识抽取模型的参数，主要是特征的权重。测试模块是基于训练部分的特征权重值在测试语料上抽取命名实体的过程。基于测试部分所抽取的命名实体，结合相应的评价指标，从而确定精确的命名实体抽取模型。具体的模型构建流程见图 3。

（2）评价指标

命名实体知识抽取的评价指标用精确率 P（Precision）、召回率 R（Recall）和调和平均值（F- Score）。具体的命名实体知识抽取的精确率和召回率计算公式如下：

$$\text{精确率 } P = \frac{\text{正确抽取实体}}{\text{正确抽取的实体+ 错误抽取的实体}} * 100\% \quad (3)$$

$$\text{召回率 } R = \frac{\text{正确抽取的实体}}{\text{正确抽取的实体+ 没有抽取的实体}} * 100\% \quad (4)$$

在用精确率和召回率来评价命名实体抽取的性能的过程中，提高召回率时，精确率会下降，反之亦然。在这种情况，采用 P 和 R 的调和平均值 F 作为综合的评价指标。具体的计算公式如下：

$$\text{调和平均值 } F = \frac{(\beta^2+1)*P*R}{(\beta^2*P)+R} = \frac{2*P*R}{P+R} \quad (\text{当}\beta=1) \quad (5)$$

4.4 命名实体的抽取

基于条件随机场，通过自身模板和添加特征模板，选取经过标注的语料进行训练和测试，从而确定命名实体知识抽取的模型。在具体的测试过程中，为了使所得结果更加合理和科学，采取了交叉验证，把训练和测试的语料按 9∶1 的比例共分成了 10 份，分

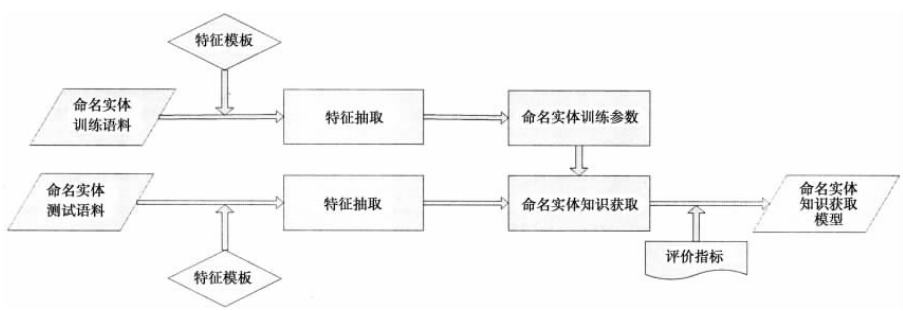


图 3 命名实体知识抽取模型构建流程

别进行训练和测试，以期从中得到最优的知识抽取模型。具体见表 6。

表 6 基于条件随机场构建的模型性能

在基于词汇、词位和词性组合的特征基础上，针对命名实体知识抽取的任务，从基于条件随机场的抽取模型的调和平均值上可以看出，本模型的精确率和召回率能都达到 80% 以上，基本能够满足对项目申请书中命名实体的抽取要求。基于条件随机场构建的抽取模型最好的 F 值为 86.22%，这说明本模型在对项目书中的命名实体的抽取中的性能是比较突出的。

5 结语

项目申请书中信息自动化抽取，对于企业项目申报具有重要意义，可以帮助服务提供单位更加有效地给企业提供知识服务。本文根据项目申请书中命名实体的分布特点，构建了基于条件随机场的（CRFs）的命名实体抽取模型，并通过实验对该模型的抽取性能进行了检查。实验结果表明，模型能够较好地对项目书中的命名实体进行抽取，基本能够满足自动抽取的要求。

参考文献

[1] 命名实体识别测评组. 2004 年度命名实体识别测评大纲[C]. 863 计划中文信息处理与智能人机交互测评会议，2004.

[2] Ralph Grishman. The NYU system for MUC-6 or where's the syntax[C]. Proceedings of the Sixth Message Understanding Conference, Morgan Kaufmann, 1996.

[3] Humphreys K, Gaizauskas R, et al. University of Sheffield: Description of the Lasie-II system as used for MUC-7[C]. Proceedings of MUC-7, Washington D. C. 1998.

[4] Krupka George R, Hausman Kevin. IsoQuestInc: Description of the NetOwl extractor system as used for MUC-7[1998][C]. Proceedings of the Seventh Message Understanding Conference (MUC-7), Washington D. C. 1998.

[5] 孙茂松，等. 中文姓名的自动辨识[J]. 中文信息学报，1995，9(2)：16-27.

[6] 吕雅娟，等. 基于分解与动态规划策略的汉语未登录词识别[J]. 中文信息学报，2001，15(1)：123-128.

[7] Daniel M Bikel, Richard Schwartz, et al. An algorithm that learns what's in a name[J]. Machine Learning, 1999, 34(1-3)：211-231.

[8] 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.

[9] Yamada Hiroyasu, Kudoh Taku, et al. Japanese named entity extraction using support vector machine[J]. IPSJ Journal, 2002, 43(1)：44-53.

[10] Kazama Jun'ichi, Takaki Makino, et al. Tuning support vector machines for biomedical named entity recognition[C]. Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, 2002.

[11] Andrew McCallum, Wei Li. Early results for named entity recognition with conditional random fields feature induction and web-enhanced lexicons[C]. Proceedings of CONLL-2003, Edmonton, Canada, 2003.

[12] Lafferty J, McCallum A, et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]. Proceedings of International Conference on Machine Learning. San Francisco: Morgan Kaufman, 2001: 282-289.

[13] 王春雨，王 芳. 基于条件随机场的农业命名实体识别研究[J]. 河北农业大学学报，2014，37(1)：132-135.

[作者简介] 王文龙，男，1989 年生，南京大学信息管理学院硕士研究生。

王东波，男，1981 年生，南京农业大学信息科学技术学院副教授。

收稿日期：2014-09-10

信息窗

第十八次全国社科院图书馆馆长协作会议在福州举行

2014 年 11 月 24 日，由福建社会科学院主办的第十八次全国社科院图书馆馆长协作会议暨地方社科院智库建设与文献信息服务论坛在福州举行。本次会议得到中国社会科学院和各兄弟省、市、自治区社科院的高度重视，34 个社科院的 90 多位代表参会。中国社科院图书馆党委书记庄前生应邀在开幕式上致辞，福建社会科学院党组书记陈祥健致欢迎辞，副院长李鸿阶主持了开幕式。

在论坛上，专家学者围绕新形势下智库建设与文献信息服务的主题展开研讨。与会代表一致认为，各社科院图书馆应加强移动化、数字化建设，以打造先进的情报收集系统和科研协同平台，为研究人员提供更加高效便捷的服务。会上，十余家社科院代表倡议建设“社会科学大数据平台”，平台的主要任务是多方合作进行大数据协同研究，发挥各地专业、地方特长，通过联盟共享交互数据，达到数据利用的合理化，各社科院图书馆代表在倡议书上签字。（资 信）