基于 SVM 和泛化模板协作的藏语人物属性抽取 *

朱臻 1,2，孙媛 1,2

（1.中央民族大学信息工程学院，北京市 100081；

2.中央民族大学国家语言资源监测与研究中心少数民族语言分中心，北京市 100081）

摘要：该文提出了一种基于 SVM 和泛化模板协作的藏语人物属性抽取方法。该方法首先构建了基于藏语语言规则的模板系统，收集了包括格助词、特殊动词等具有明显语义信息的特征建设模板并泛化。针对规则方法的局限性，该文在模板的基础上，采用 SVM 机器学习方法，设计了一种处理多分类问题的层次分类器结构，同时对多样化的特征选取给予说明。最后，实验结果表明，基于 SVM 和模板相结合的方式可以对人物属性抽取的性能有较大提高。

关键词：人物属性抽取；藏语语言处理；SVM；层次分类器

中图分类号：TP391　　　　**文献标识码：**A

Tibetan Person Attributes Extraction Based on SVM and Pattern

Zhen Zhu[1,2], Yuan Sun[1,2]

(1.School of Information Engineering, Minzu University of China， Beijing, 100081, China;

2. Minority Languages Branch, National Language Resource and Monitoring Research Center, Beijing, 100081, China)

Abstract: This paper proposes an SVM and pattern based approach to Tibetan person attributes extraction. Starting from the pattern system built with language rules. In which, the Tibetan language features with clear semantic information are vital, such as case-auxiliary words, particular verb and etc., and then normalization by some experimental results. Secondly, considering the shortage of rule method, machine learning approach are introduced, the SVMs with multiply feature vectors are used and organized by a hierarchy classifier strategy. Finally, experiment results prove this method has greater improvement in person attributes extraction.

Keywords: Person attributes extraction; Tibetan language processing; SVM; Hierarchy classifier

1 引言

随着互联网的快速普及，特别是发展中国家互联网用户的快速增加，网络上非英语文本资源数量急速增长，其增长速度远远超过了 10 年前的速度，并且越来越多的网上信息以多语言的形式发布。

据中央民族大学国家语言资源监测与研究中心少数民族语言分中心调查：截止到 2013 年 12 月底，大陆少数民族语言文字的网站总量在 1,250 个左右，其中维吾尔文网站 840 个、藏文网站 146 个、蒙古文网站 136 个。“与全国网民增长速度相比，少数民族网民的增速较为突出，例如藏族网民增幅达 86%，远远高于全国平均增长速度”[1]。

Web 内容的爆炸式增长，使得对 Web 的社会网络研究已经不再局限于对 Web 结构的分析，而是转向以 Web 内容为研究对象的分析[2]，其中知识图谱（Knowledge Graph）成为大数据时代自然语言处理领域的一个研究热点。知识图谱以结点代表实体或者概念，边代表实体/概念之间的各种语义关系，其中实体知识的抽取是主要研究内容之一。

知识图谱以全面、完整的知识体系为信息检索、问答系统、知识库构建等领域的研究提供了资源和支撑，而目前已有的 Google（超过 5.7 亿实体，18 亿关系），DBpedia（超过 1900 万实体，1 亿关系），Wiki-links（4000 万排除歧义的关系），Wolframalpha（10 万亿关系），

---

* **收稿日期：**2015-06-15　**定稿日期：**2015-08-10

基金项目：国家自然科学基金项目（No.61331013）；北京青年英才资助计划（No.YETP1291）；国家语委项目（No.ZDI125-36，No.YB125-139）；中央民族大学自主科研项目（No.2015MDQN11）；中央民族大学国家语言资源监测与研究中心少数民族语言分中心项目（No.CML15B02）

Probase（超过 265 万实体），百度知心，搜狗知立方等知识图谱只提供英、汉、法等语言的相关知识[3]，少数民族语言知识图谱的构建才刚刚起步。

例如，当搜索“ཏཱ་ལའི་བླ་མ་（达赖喇嘛）”时，Google 会出现 64,100 条结果；而当搜索“རྒྱལ་བ་རིན་པོ་ཆེ་（嘉瓦仁波切）”时，Google 会出现 586,000 条结果。在藏语中，通常称 ཏཱ་ལའི་བླ་མ་（达赖喇嘛）为 རྒྱལ་བ་རིན་པོ་ཆེ་（嘉瓦仁波切），而目前的搜索引擎却没有显示两者之间的关系。此外，所有搜索结果以含有关键词的文本显示为主，没有知识的结构表示。如果具有了实体与实体之间的语义链接，有了实体知识，那么将会获得更全面的信息，实现信息的深度挖掘。

因此，本文针对藏语语言的特点，提出了一种基于 SVM 和泛化模板协作的藏语人物属性抽取方法。藏语人物属性抽取的研究，是藏语知识图谱构建的基础，为藏语知识问答、信息检索、信息抽取等领域研究提供支撑，对提高少数民族地区的社会管理科学化水平、维护民族团结和国家统一、构建和谐社会具有重要意义。

2 国内外研究现状及发展动态分析

人物属性抽取是信息抽取领域的一个重要领域[4]，该概念在 2009 年的国际 TAC KBP 会议开始引入[5]。人物属性抽取是指自动从无结构或者半结构的文本语料中抽取特定的人物属性，其中包括人物性别，出生年月，出生地，工作地点等。但是人物属性抽取一直面临着两大问题[6]，即人物属性识别问题和人物属性关系判别问题。人物属性一般为命名实体，例如人名、地名和组织机构名。命名实体识别在自然语言处理领域仍是一件尚未完全解决的工作。因此，在人物属性抽取工作前，需要准备高准确度命名实体标注语料[7]。

为了实现大规模数据的信息抽取，很多机器学习算法被引入到信息抽取领域。Freitag 采用 HMM 结构进行信息抽取[8]，Laffery 使用条件随机场抽取数据[9]，Kambhatla 把多种特征用于最大熵模型并取得了较好的抽取效果[10]。而应用最广的是支持向量机方法[11][12]。作为信息抽取领域的一个分支，把统计的方法运用于人物属性抽取，通常采用基于特征向量的方式[13]。其中，经典的基于特征向量的机器学习方法包括最大熵模型[14]和支持向量机[15]。另外，特征选取对于基于特征向量的方式至关重要。Miler 构建了一种语义解析树，树中整合了概念间关系的多种语义信息，包括词性标注，命名实体识别标记和其他一些语言上的强特征，这些特征给分类器提供了很好的依据[16]。Culotta 根据依存树构建了核函数，并将其用于机器学习算法[17]。Zelenko 引入了一种树核的方法[18]。

但是，目前对于藏语的实体知识抽取领域的研究较少，主要研究集中于藏语的命名实体识别方法[19-21]，而对于实体关系抽取特别是人物属性抽取的研究尚未有成熟的成果。归纳原因，藏语任务属性抽取存在的困难如下：（1）训练语料匮乏；（2）藏语在句子和篇章级的信息处理研究还处于起步阶段，因此，英、汉实体关系抽取中的核函数方法无法直接应用于藏语实体关系抽取中。

因此，本文针对藏语的特点，构建了一定规模的训练语料，提出一种基于 SVM 和泛化模板的藏语人物属性关系抽取方法。其中，模板构建重点选取包括藏语后置谓词，相关的格信息等主要特征。此外，针对模板方式的局限性，本文采用 SVM 机器学习方法，设计了一种处理多分类问题的层次分类器进行属性关系抽取。最后，本文分别采用模板、SVM 以及模板和 SVM 结合的方法进行实验，实验结果表明，通过模板和 SVM 结合的方式有效提高了人物属性抽取的正确性。

3 整体框架

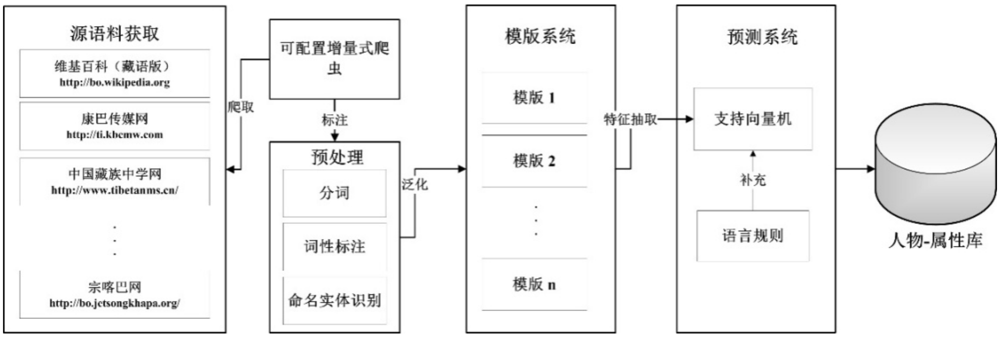

图 1 基于 SVM 和泛化模板相结合的藏语人物属性关系抽取方法

通过可配置的爬虫系统从多个藏文网站获取语料，从中筛选出关于人物介绍的文章并对这些句子做预处理，包括分词，词性标注和命名实体识别。首先，根据训练语料构建模板系统。此外，为了应对开放语料的多样性问题，引入了 SVM 方法进行预测，而模板中的语言规则作为辅助工具。最终将处理完成的数据装入人物-属性库中。

4 模板构建

一定量的训练语料标注之后，可以进入模板系统建设阶段，本章将分别介绍藏语特征选择，模板建设和泛化过程。

4.1 主要藏语特征选择

不同于汉语和英语，藏语是谓语后置型语言，动词是句子的核心。动词附近的格标记含有丰富的语义角色信息，格标记在一定程度上反映出句子中谓词与主体词之间的关系，而且这些格标记的出现存在一定的规律。因此，对格标记做了整理，这些格标记对藏文人物属性抽取起到重要的作用，如表 1 所示。

表 1 藏语格标记的类型与作用

类型	格标记的列举	包含类型	语法、语义作用
主格	གིས、ཀྱིས、གྱིས、ཡིས、འིས	施格	指明动作的施动者
		工具格	指明动作的工具、方式等
属格	གི、གྱི、ཀྱི、འི、ཡི		表示领属关系
拉格	སུ、རུ、ར、ལ、ཏུ、དུ、ན	业格	指明对象、地点等
		为格	表示受益的对象和动作
		依格	表示依存或所在等
		同体格	表示事物的状态
		时间格	表示发生的时间
从格	ནས、ལས		表示动作或状态的来源

4.2 模板建设

与汉语和英语不同，模板建设中更加注重藏语特有的格标记和周围的动词，在语料标注的基础上构建特征模板，如例 1-4。

例 1：ཚེ་བརྟན་སྒྲོལ་མ་ནི་བོད་ལྗོངས་གཞིས་ཀ་རྩེ་རུ་སྐྱེ་འཁྲུངས།（才旦卓玛出生在西藏日喀则。）

模板：<人名/nh>(ནི/v)<地点/ns>(རུ་/k)(སྐྱེས་ 出生/v)

例 2：བསྟན་འཛིན་རྒྱ་མཚོའི་ཡབ་ཆོས་སྐྱོང་ཚེ་རིང་ཡིན།（旦增加措的父亲是其将才让。）

模板：<人名/nh>(འི/k)(ཡབ་ 父亲/n)(ནི་ 是/v)<人名/nh>

例 3：ཚེམཁན་པོ་བསོད་ནམས་དར་རྒྱས་སྤྱི་ལོ་༡༩༦༢ལོར་སྐྱེ་འཁྲུངས།（堪布索南达吉出生于公元 1962 年。）

模板：<人名/nh><时间/t>(ར་/k)(སྐྱེ་འཁྲུངས 出生于/v)

例 4：པཎ་ཆེན་སྐུ་ཕྲེང་བརྒྱད་པ་བསྟན་པའི་དབང་ཕྱུག་གི་ཡུམ་བཀྲ་ཤིས་ལྷ་མོ།（班禅第八世丹巴昂秀的妈妈是扎西拉姆。）

模板：<人名/nh>(གི་/k)(ཡུམ་ 妈妈/n)(ནི་ 是/v)<人名/nh>

词性标记采用“国家语言资源监测与研究中心少数民族语言分中心”的《信息处理用现代藏语词类标记集规范》，其中，“/nh”表示人名、“/t”表示时间、“/ns”表示地名、“/k”表示格标记、“/v”表示动词。

4.3 泛化

在语料模板建设完成后，发现众多模板具有相似性，我们整合、修改并泛化模板使其能应用于更广泛的语料。对于微小区别模板，例如仅是动词的差别，只需将不同的动词添加的集合来合并模板。对于模板中不重要的修饰性成分，将其从模板中删除，模板样式如例 5-8。

例 5：ཚེ་བརྟན་སྒྲོལ་མ་ནི་བོད་ལྗོངས་གཞིས་ཀ་རྩེ་རུ་སྐུ་འཁྲུངས།（才旦卓玛出生在西藏日喀则。）

模板：<人名/nh>(ནི/v)<地点/ns>(སུ་|ར་|རུ་|དུ་|ཏུ་|ལ་|ན་/k)(སྐྱེས་|འཁྲུངས 出生/v)

例 6：སྒྲོལ་དཀར་གྱི་ཕ་ཡུལ་ནི་མཚོ་བོད་དུ་ཡིན།（卓嘎的家乡在青海。）

模板：<人名/nh>(གི|གྱི|ཀྱི|འི|ཡི/k)(ཕ་ཡུལ་ 家乡/n)(ནི་ 指示词/r)<地名/ns>(སུ|རུ|ར|ལ|ཏུ|དུ|ན/k)(ཡིན། 是/v)

例 7：ཚེམཁན་པོ་བསོད་ནམས་དར་རྒྱས་སྤྱི་ལོ༡༩༦༢ལོར་སྐུ་འཁྲུངས།（堪布索南达吉出生于公元 1962 年。）

模板：<人名/nh><时间/t>(སུ་|ར་|རུ་|དུ་|ཏུ་|ལ་|ན་/k)(སྐུ་འཁྲུངས 出生于/v)

例 8：ཚེ་དབང་གི་སྐྱེས་སྐར་ནི་1988ལོའི་ཟླ10བའི་ཚེས1ཉིན་ཡིན།（次旺的生日是 1988 年 10 月 1 日。）

模板：<人名/nh>(གི|གྱི|ཀྱི|འི|ཡི/k)(སྐྱེས་སྐར་ 生日/n)(ནི་ 指示词/r)<时间/t>(ཡིན 是/v)

5 基于 SVM 的层次分类

虽然基于特征模板的方法在特定的测试语料中可以取得较高的准确度，但是它需要很多人工的介入并且对于模板系统尚未覆盖的内容无能为力。因此，对于不同的语料准确率和召回率差别很大，特别是对于模板系统比较生疏的语料，基于模板的抽取系统召回率非常低。因此，引入了基于特征向量的 SVM 方法，并设计了层次分类器。

5.1 特征选取

特征选择至关重要。一定程度上，特征的质量决定了分类效果。本文的特征向量主要选取关键词特征，标注组合特征，实体词周围标记特征。

5.1.1 关键词特征

关键词指出现频率较高并且含有极强区分特性的名词或动词。这些特征大多是从模板系统提取出来的，虽然关键词特征向量数量并不多，但是这些词往往具有很强的区分度并且这些特征会以高的频率出现在某一属性类别中，例如，关键词名词ཡུམ་|ཨ་མ་（妈妈|母亲）。

5.1.2 基于多种标记的组合特征

相比于基于词本身的特征，基于词性标注的特征更具有广泛性。但是不是每个标记都可以作为特征向量，因为众多标记并没有区分度。因此，本文主要采用标记组合特征，特别是格标记和词性或命名实体标记组合往往能起到较好的分类效果。例如，时间标记“/t”+格标记“/k”+/v（如སྐུ་འཁྲུངས 出生）对于识别出生年月属性有较大的帮助。

5.1.3 实体词周围标记特征

实体词周围标记特征是指在实体词周围的词标记构成的特征，包括词性标记和命名实体标记。本文认为离实体词越近的标记越重要，而离实体词距离越远的标记则较不重要。因此，选取实体词向前 2 个词距和向后 1 个词距内的词性标注标记和前后 3 个词距内的命名实体标记。

5.2 构造层次分类器

SVM 目前是信息抽取领域应用较为成功的分类器之一。SVM 通过在高维空间上寻找最优超平面，从而达到分类目的。对于非线性可分的样本集，一般是通过升维实现样本空间映射，从而转变成线性可分的问题。为了使问题可计算，即避免出现维度灾难问题，引入了核函数的方法，从而达到把计算在低维空间完成的目的。对于人物属性抽取问题，一个关键问题是构建高性能的 SVM 分类器。SVM 最初被设计用来解决二分类问题，但是属性抽取往往都是复

杂的多分类问题。例如，人物属性可以分为出生年月，出生地，性别等多个类别。那么，如何组织这些分类器则是多分类问题必须解决的问题。

目前主流的分类器组织形式分为两种：

（1）一对多的方式。假如一共有 k 个属性类别，那么需要构建 k 个分类器，并且对于每个属性确定平均需要进行 $k/2$ 次预测，此方式分类效果欠佳。

（2）一对一的方式。同样如果存在 k 个属性类别，那么需要构建 $k(k-1)/2$ 个分类器，然后通过 $k(k-1)/2$ 次预测，再计算累加权重，获得累加值最大的类别则为所属类别。这种方式比前者好，但是分类器数量过多，对于属性抽取等类别数量较多的问题适用性较差。

因此，本文引入了一种层次分类器的构造方法。该方法结合两种传统方法的长处，同一层面采用一对一的方式，逐层向下。同时，利用模板系统中获取的语言规律建设快速通道，从而进一步优化层次分类器的分类效果和分类速度。具体构造如图 2 所示。

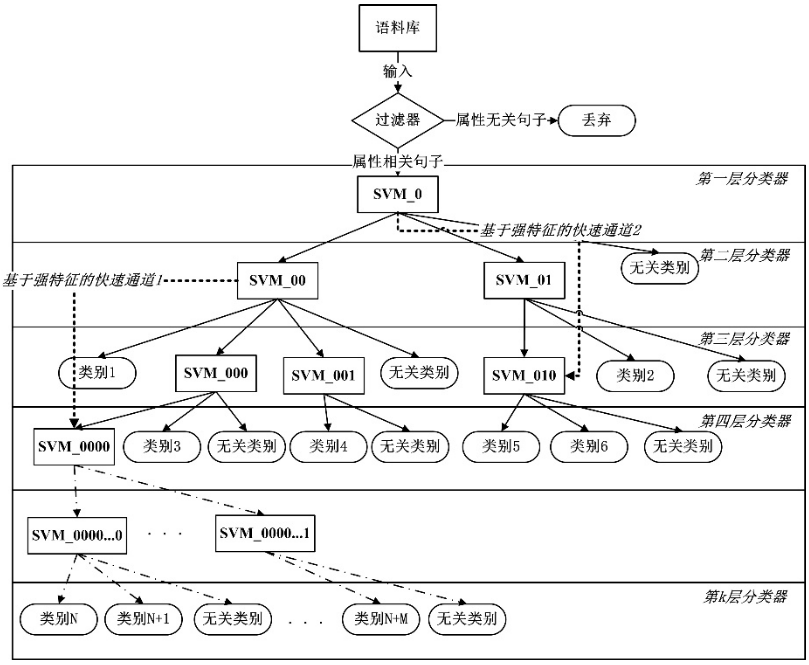

图 2 层次分类器的构造

（1）过滤器：在进入层次分类器系统之前需要对语料做筛选，将没有任何属性实体存在的部分干扰句直接剔除，可以一定程度上减少层次分类器工作负荷从而提高效率。

（2）逐层向下：进入层次分类器系统后，标准的分类模式是从第一层分类器开始逐层向下直至类别叶节点，中间的分类器会将一些无关类别的数据剔除。这一步骤对于属性抽取过程中大量负样本的处理是非常重要的。

（3）同层多分类问题：对于单个分类器，采用一对一的方式处理多分类问题。经过层次分类后，每个多分类处理的类别都不会太多，所以不会出现传统一对一问题分类器数量过大的问题。分类器的个数为 $N_{sum}=\sum_{i=1}^{n} p_i(p_i-1)/2$，其中，$p_i$ 为每个分类器中类别的个数。同时这样做可以保留一对一分类器的分类准确性高的特点。

（4）快速通道：本文设计了根据实体-属性标注构造的快速通道，这些快速通道可以有效的提高层次分类器的分类效果和速度。因为在属性抽取任务中，属性实体本身往往带有明显的区分性。例如，当出现时间为第二个实体词时，只可能出现出生年月属性而不会是父亲或出生地的属性。因此可以通过快速通道直接跳至关于出生年月类别和无关类别的分类决策器。

6 实验结果与分析

6.1 语料来源

6.1.1 数据爬取及筛选

本文语料来源于 7 家藏语网站，如表 2 所示。我们研究关注的人物属性主要包括：

人名—出生日期　　　人名—出生地

人名—父亲　　　　　人名—母亲

表 2 语料来源

语料来源	网站网址
维基百科（藏语版）	http://bo.wikipedia.org
中国藏族中学网	http://www.tibetanms.cn/
康巴传媒网	http://ti.kbcmw.com
喜马拉雅苯教（藏语版）	http://old.himalayabon.com/
AMDO 藏语	http://www.amdotibet.cn/
HIMALAYABON	http://www.himalayabon.com/
宗喀巴网	http://bo.jetsongkhapa.org/

我们从大量网页文本中选取 2400 句包含人物属性的句子。其中，1975 句是包含上述 4 种人物属性关系的句子，剩余 425 句为其他人物属性关系的句子。我们将 1600 句作为训练语料，其余 800 句作为测试语料。

6.1.2 语料预处理

我们对选取的 2400 句进行分词、词性、命名实体识别，并标注了实体之间的关系。

<e1>ཚེ་བརྟན་སྒྲོལ་མ་/nh</e1>ནི་/v<e2>བོད་ལྗོངས་གཞིས་ཀ་རྩེ/ns</e2>ཏུ་/kསྐྱེ་འཁྲུངས/v།/w

人物-出生地（e1,e2）

<e1>མཁན་པོ་བསོད་ནམས་དར་རྒྱས་/nh</e1><e2>སྤྱི་ལོ་༡༩༤༢ལོ་/t</e2>ར་/kསྐྱེ་འཁྲུངས་/v</e2>།/w

人物-出生年月（e1,e2）

6.2 实验分析与评价

首先使用基于模板的方法在 1600 句训练语料集上做测试（共包含 1705 个属性），实验结果如表 3 所示。

表 3 基于模板的藏语人物属性抽取在封闭训练集上的结果

属性类别	数量（个）			百分比（%）		
	Total	*Identified*	*Correct*	*P*	*R*	*F1*
出生年月	452	432	403	93.29%	89.16%	91.18%
出生地	458	443	407	91.87%	88.86%	90.34%
父亲	363	359	331	92.20%	91.18%	91.69%
母亲	432	425	401	94.35%	92.82%	93.58%

但是，把这些模板应用在 800 句测试语料集（共 846 个属性）时，实验结果如表 4 所示。

表 4 基于模板的藏语人物属性抽取在开放测试集上的结果

属性类别	数量（个）			百分比（%）		
	Total	*Identified*	*Correct*	*P*	*R*	*F1*
出生年月	219	162	91	56.17%	41.55%	47.77%
出生地	223	168	78	46.43%	34.98%	39.90%
父亲	184	144	73	50.69%	39.67%	44.51%
母亲	220	171	87	50.88%	39.55%	44.50%

上述实验结果表明，基于模板的方法应用在模板系统不熟悉的语料中性能下降明显。主

要原因在于，基于模板的方式缺少学习能力而必须通过一些人工参与构建，虽然通过不停的泛化和修正，性能会逐渐提升，但是过多的人工介入和较大的工作量成为该方法的瓶颈。此外，不同藏语地区或不同风格的网站的语言会有一些区别，考虑语言的丰富性，难以通过基于模板的方式做到完备。

下面，我们采用基于 SVM 的层次分类器进行人物属性抽取，本文采用层次分类器在分类速度上较之一对一的分类器有较大提升，而两种方法的准确性相差不大。并通过语言规则构建的快速通道使分类性能更好。在实验中，我们对比常见的核函数方法，最终选型为 RBF（径向基函数）并设置参数$\gamma = 1/k$， k 为类别个数。同时考虑到语料普遍存在不均衡性，负样本大大多于正样本，因此，对正负样本分别设置了不同的惩罚因子C_+和C_-。其中， C_-为 3，正样本满足$C_+ = (Num_- / Num_+) \times C_-$。其中 Num_- 为负样本数， Num_+ 为正样本数，我们通过增大正样本的惩罚因子，从而减少因为数据倾斜造成的影响。实验结果如表 5 所示。

表 5 基于 SVM 的藏语人物属性抽取在开放测试集上的结果

属性类别	**数量（个）**			**百分比（%）**		
	Total	***Identified***	***Correct***	***P***	***R***	***F1***
出生年月	219	202	103	50.99%	47.03%	48.93%
出生地	223	211	94	44.55%	42.15%	43.32%
父亲	184	176	83	47.16%	45.11%	46.11%
母亲	220	208	101	48.56%	45.91%	47.20%

实验结果表明，相比于模板的方法，SVM 方法提高了人物属性抽取的召回率，但是准确率并没有提高。主要原因在于，SVM 的结果在对于一些不明显的分类，通过多样化的特征向量反而可以取到较好的预测效果。但是对于一些非常明显的分类问题却判断错误，我们认为，部分原因在于训练语料不足和训练语料不均匀造成的。

最后，本文采用基于模板和 SVM 相结合的方式进行实验。实验结果如表 6 所示。

表 6 基于 SVM 和泛化模板协作的藏语人物属性抽取在开放测试集上的结果

属性类别	**数量（个）**			**百分比（%）**		
	Total	***Identified***	***Correct***	***P***	***R***	***F1***
出生年月	219	201	131	65.17%	59.82%	62.38%
出生地	223	209	133	63.64%	59.64%	61.57%
父亲	184	161	108	67.08%	58.70%	62.61%
母亲	220	201	128	63.68%	58.18%	60.81%

首先对前期建设的模板系统精心筛选，只保存在抽取实验中准确率接近 100%的这部分模板。虽然这样会使召回率在模板系统部分急剧下降，但是，随后我们就将所有模板没有抽取出属性所剩下的所有句子数据化并交给 SVM 预测。这样，对于那些模板并未抽取的属性可以通过 SVM 预测出，保护了一些原本特征明显的属性句子不被 SVM 误判。所以在整体上并未影响召回率，同时还提高了抽取的效果。

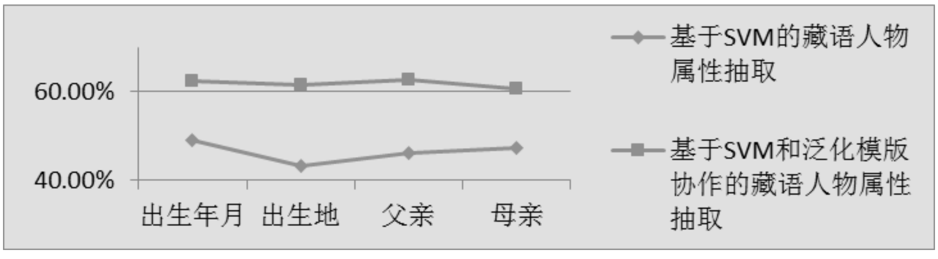

图 3 基于 SVM 和泛化模板协作和基于模板的藏语人物属性抽取比较

由图 3 可以看出，采用模板和 SVM 相结合的方式比只采用 SVM 的方式，性能上有较大的提高。

6.3 实验结果的展示

通过 SVM 和模板结合的人物属性抽取后的结果如表 7 所示。通过人物属性抽取，把属性放入人物-属性库中，为藏语人物收集，藏语知识图谱建设等应用提供数据支撑。

表 7 关于人物 སྲོང་བཙན་སྒམ་པོ་（松赞干布）的属性抽取结果

属性类别	抽取属性值	属性所在句子
出生地	གྲོང་ཁྱེར་ལྷ་སའི་ཤར་ཕྱོགས་སའི་སུ་རྒྱང་ཐག་ལི་དབར་བཅུ་ཕྲག་ཙམ་ཡོད་པའི་དབུ་ཏུ་མལ་གྲོ་རྒྱ་མ་ཕོ་བྲང་བྱམས་པ་མི་འགྱུར་གླིང་།	སྲོང་བཙན་སྒམ་པོ་ནི་གྲོང་ཁྱེར་ལྷ་སའི་ཤར་ཕྱོགས་སའི་སུ་རྒྱང་ཐག་ལི་དབར་བཅུ་ཕྲག་ཙམ་ཡོད་པའི་དབུ་ཏུ་མལ་གྲོ་རྒྱ་མ་ཕོ་བྲང་བྱམས་པ་མི་འགྱུར་གླིང་ནས་འཁྲུངས།
出生年月	བོད་མེ་གླང་（༦༡༧）ལོར་འཁྲུངས་	སྲོང་བཙན་སྒམ་པོ་ནི་བོད་མེ་གླང་（༦༡༧）ལོར་འཁྲུངས་
父亲	རྗེ་གནམ་རི་སྲོང་བཙན་	སྲོང་བཙན་སྒམ་པོའི་ཡབ་ནི་རྗེ་གནམ་རི་སྲོང་བཙན་ཡིན།
母亲	འབྲི་བཟའ་ཐོད་དཀར་ཚེ་སྤོང་བཟའ་	སྲོང་བཙན་སྒམ་པོའི་ཡུམ་ནི་འབྲི་བཟའ་ཐོད་དཀར་ཚེ་སྤོང་བཟའ་ཡིན།

7 结论

通过对上述实验结果的分析,发现对于人物属性关系抽取的问题采用 SVM 和模板相结合的方式,比仅采用 SVM 或者仅采用模板的方式性能更好。部分原因在于彼此对于不同情况的分类问题具有各自的优势,通过整合两者方法,让它们协同工作,从而使实验方法性能提高。通过该方法提取的属性可以广泛应用于专门数据库的建设、知识图谱构建和智能问答等领域。在将来的工作中，需要扩充语料库并增加人物属性的类别，从而提升成果的价值。

参考文献

[1] 李光,钟雅琼.大陆研拟藏维文网络舆情监测系统监控分裂风险[J].凤凰周刊, 2012(18).

[2] Bizer C，Heath T，Berners-Lee T. Linked data-the story so far [J].International Journal on Semantic Web and Information Systems (IJSWIS)，2009，5(3): 1-22.

[3] 张静,唐杰.下一代搜索引擎的焦点：知识图谱[J].中国计算机学会通讯, 2012,9(4):64-68.

[4] Kong Fang, Zhou Guodong, Zhu Qiaoming. Survey on Coreference Resolution [J]. Computer Engineering, 2010, 36(8): 33-36.

[5] Bikel D., Castelli V., Florian R. Entity linking and slot filling through statistical processing and inference rules[A]. In Proc. TAC 2009 Workshop[C], November 2009.

[6] Burman, A., Jayapal, A., Kannan, S.Entity linking, slot filling and temporal bounding[A].KBP[C] 2011.

[7] Axel Bernal, Koby Crammer, Artemis Hatzigeorgiou. Global discriminative learning for higher-accuracy computational gene prediction[J] PLoS Computational Biology, 2007, 3(3).

[8] Freitag D., and McCallum A. Information extraction with HMM structures learned by stochastic optimization[A].AAAI Press[C], Menlo Park, CA: 2000, 584-589.

[9] Lafferty, J., McCallum, A., Pereira. F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[A]. In Proc. 18th International Conf. on Machine Learning[C], Morgan Kaufmann, San Francisco, CA: 2001, 282-289

[10] Kambhatla N. Combining lexical, syntactic and semantic features with Maximum Entropy models for extracting relations[A]. Proceedings of 42th Annual Meeting of the Association for Computational Linguistic[C], July Barcelona, Spain: 2004, 21-26.

[11] Zhou G., Su, J., Zhang, J., Zhang, M. Combining Various Knowledge in Relation Extraction[A].Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics[C],2005.

[12] Zelenko D., Aone C., Richardella. Kernel methods for relation extraction[J]. Journal of Machine Learning Research, 2003, 1083-1106.

[13] Nadia Ghamrawi and Andrew McCallum. Collective multi-label classification[A]. In Conference on Information and Knowledge Management (CIKM)[C], 2005.

[14] Nanda Kambhatla. Combining lexical, syntactic and semantic features with Maximum Entropy models for extracting relations[A]. Proceedings of ACL[C], 2004, 178-181.

[15] Zhao S B, Grishman R. Extracting relations with integrated information using kernel methods [A]. Proceedings of ACL[C], 2005, 419-426.

[16] Miller S., Fox H., Ramshaw L. and Weischedel R. A novel use of statistical parsing to extract information from text [A]. In Proceedings of 6th Applied Natural Language Processing Conference[C], Seattle, USA. 2000.

[17] Culotta A. and Sorensen J. Dependency tree kernels for relation extraction[A]. In Proceedings of 42th Annual Meeting of the Association for Computational Linguistics[C], Barcelona, Spain: July 2004, 21-26.

[18] Zelenko D., Aone C. and Richardella. Kernel methods for relation extraction[J]. Journal of Machine Learning Research, 2003, 1083-1106.

[19] 加羊吉，李亚超，宗成庆，于洪志. 最大熵和条件随机场模型相融合的藏文人名识别方法 [J]. 中文信息学报，2013.

[20] 才智杰. 藏文自动分词系统中紧缩词的识别 [J]. 中文信息学报，2009，23(1): 35-37.

[21] Sun Yuan, Zhao Xiaobing. Research on automatic recognition of Tibetan personal names based on multi-features [A]. Proceedings of International Conference on Natural Language Processing and Knowledge Engineering[C], 2010.

作者简介：

朱臻（1988——），男，硕士研究生，主要研究领域为自然语言处理、信息检索、数据挖掘。
E mail:19057736389@163.com

孙媛（1979——），通信作者，女，副教授，中文信息学会会员，主要研究领域为自然语言处理、信息抽取。
Email:tracy.yuan.sun@gmail.com。