Modern Linguistics 现代语言学, 2018, 6(3), 449-461
Published Online August 2018 in Hans. http://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2018.63051

Hans汉斯

A Survey of Lao Named Entity Recognition

Yangyu He, Mianzhu Yi, Huixin Jia, Hongxin Li*

PLA Strategic Support Force Information Engineering University, Luoyang Henan
Email: *lihongxin830@163.com

Received: Jul. 22nd, 2018; accepted: Aug. 1st, 2018; published: Aug. 7th, 2018

Abstract

Named entity recognition is a key basic technology in natural language processing and has become an important part of many tasks such as information retrieval, machine translation, and so on. The recognition of Lao named entity has begun to attract attention in recent years, and the similarity of Lao and Chinese makes dealing with NER a challenge. This article briefly reviews the research history of domestic and foreign named entity recognition, introduces the latest achievements of Lao word segmentation and named entity recognition, and compares the advantages and disadvantages of each method. Finally, it forecasts the development trend of Lao named entity recognition.

Keywords

Lao Language, Named Entity Recognition, Word Segmentation, Information Retrieval, Information Extraction

老挝语命名实体识别研究综述

何阳宇，易绵竹，贾惠心，李宏欣*

中国人民解放军战略支援部队信息工程大学，河南 洛阳
Email: *lihongxin830@163.com

收稿日期：2018年7月22日；录用日期：2018年8月1日；发布日期：2018年8月7日

摘　要

命名实体识别是自然语言处理中一项关键的基础技术，已成为信息检索、机器翻译等诸多任务的重要组成部分。在诸多语种的命名实体识别研究工作中，老挝语命名实体识别近年来开始受到关注，然而老挝语与汉语类似的特点又使得其命名实体识别工作非常具有挑战性。本文对国内外命名实体识别的研究历

*通讯作者。

文章引用: 何阳宇, 易绵竹, 贾惠心, 李宏欣. 老挝语命名实体识别研究综述[J]. 现代语言学, 2018, 6(3): 449-461.
DOI: 10.12677/ml.2018.63051

程进行了简要回顾，同时介绍了老挝语分词和命名实体识别的最新成果，并对比分析了每种方法的优点和不足。最后，对老挝语命名实体识别的发展趋势作了展望。

关键词

老挝语，命名实体识别，分词，信息检索，信息抽取

1. 引言

最早研究命名实体的是 Lisa F. Rau (1991)，他在第七届 IEEE 人工智能应用会议上介绍了依靠启发式算法和手工制定的规则“识别和抽取公司名称”的系统，但当时没有引起太大关注[1]。现在广泛使用的“命名实体(Named Entity)”这个术语是在第六届信息理解会议(MUC-6)上首次使用的[2]，在这次会议上，NERC (Named Entity Recognition and Classification)也作为信息抽取的子任务而被提出，其目的是从报纸等非结构化文本中自动提取出结构化文本(如专有名词等)。自 MUC-6 之后，人们对人名、地名和组织机构名三类实体的识别粒度进一步细化，如地名可进一步划分为城市名、州名和国家名，这就形成了一个实体层级结构。根据特定的需要，还引入了其他类型的实体，如电子邮件地址、电话号码、书名等。

早期的命名实体识别大多是基于规则的。Ralph Grishman 于 1995 年开发了一套 NER 系统，该系统使用一些专业名词词典，包括所有国家名、主要城市名、公司的名称以及常用人名等。1999 年，Borthwick 开发了基于机器学习(最大熵模型)的命名实体识别系统，该系统使用了 8 部词典。MUC-7 的八个 NER 参赛系统中有 5 个是基于规则的，CoNLL-2003 展示了 16 个基于规则的 NER 系统。20 世纪 90 年代末，尤其是进入 21 世纪以后，这种单纯依靠规则来进行命名实体识别的方法遇到越来越多的问题：NER 系统只识别存储在词典中的命名实体，这种识别方法虽然简单快速，但是命名实体数量太多，并且在不断发展，即使是已经列入词典的命名实体可能也存在歧义。例如，“华盛顿”既指人名，也指地名。于是，基于大规模语料库的统计方法开始流行。该方法将命名实体识别问题视为序列标记问题，这就使得命名实体识别相对独立于语言领域，不需要太多语言学专业知识，可移植性也大为增强。基于统计方法的命名实体识别大致可分为无监督、半监督、有监督三种学习方法以及混合方法。无监督学习方法不需要标注训练语料(Shinyama, Sekine, 2004)，半监督学习方法只需少量的已标注语料(Collins 和 Singer, 1999)。当有大量的高质量标注语料可用时，可采用有监督学习方法中的隐马尔可夫模型(Bikel 等，1997)、决策树(Sekine 等，1998)、最大熵模型(Mikheev 等，1998)、支持向量机(Asahara & Matsumoto, 2003)和条件随机场(McCallum, 2003)等模型的良好性能来得到最优化结果[3]。

目前，大多数命名实体识别研究集中在英语上，方法设计也主要考虑英语的语言特点。汉语命名实体识别研究虽然起步较英语稍晚，但也取得了较为丰硕的成果。相比之下，老挝语在该领域的研究非常匮乏，主要原因是老挝经济社会发展较为落后，科技水平和科研能力较弱。但随着互联网时代和信息时代的到来，老挝社会也逐渐跟上了世界的步伐，越来越多的老挝语信息通过网络实现传播，包括中国在内的世界各国与老挝的往来也日益密切，因此，能够从海量的老挝语信息中提取重要实体成为了新的迫切需要。

老挝语属于汉藏语系壮侗语族壮傣语支，是一种孤立型、分析型语言，没有丰富的形态变化。除了老挝，泰国东北部和北部的老挝族也使用老挝语。老挝文字属于表音文字中的音位文字类型，有两种不同的形体。一种较古老的称为“多坦(ຕົ໋ວທຳ)”，意为经文，多见于贝叶经中，在佛教寺庙中使用；另一种称为“多老(ຕົ໋ວລາວ)”，意为老文，形状和拼写都类似现代的泰文[4]。

本文主要考察国内外老挝语命名实体识别的研究进展，分析老挝语在该领域面临的困难和挑战，以及展望未来的发展方向。

2. 老挝语命名实体识别相关研究成果分析

在自然语言处理中，分词和命名实体识别有着密不可分的关系。分词系统中加入命名实体识别模块，或者将命名实体作为辅助信息加入分词系统，可以提高分词系统性能。同样，分词效果的提升也能够为命名实体识别提供帮助。老挝语跟汉语一样，都是孤立语，词与词之间没有空格等显性标志来指示词边界，汉语中许多实验研究都表明，集外词是造成分词错误的主要原因，其中超过一半的错误是由于命名实体造成的。因此，本文将会把这两方面的主要研究成果同时展现出来。

2.1. 老挝语分词

2.1.1. 老挝语分词中的主要问题

老挝语分词的基本问题大致分为三个方面：分词规范、歧义切分和未登录词识别。老挝语学界尚未明确“词”的定义，在实际操作中，人们很难界定“词”和“词组”；歧义切分大致有交集型和组合型两种，处理这类问题需要考虑上下文信息以及韵律；未登录词又称为生词(unknown word)或者集外词(out of vocabulary, OOV)，包括词表中未收录的词以及训练集以外的词，对于大规模语料来说，未登录词对分词精度的影响要远高于歧义切分[5]。老挝语中包含大量的未登录词，一方面是由于老挝语词典资源较少，另一方面是因为随着网络的普及，大量外来词涌入老挝语，而老挝对于如何翻译这些外来词没有统一的规定，这就对分词造成了极大的困难。

2.1.2. 老挝语分词的主要方法

总的来说，老挝语分词方法可分为两种不同的类型：基于词表(DCB)和基于机器学习(MLB)或者说基于统计[6]。

A. 基于词表的方法

基于词表的方法主要依赖词表中的一系列词语来解析输入文本并将其切分成单词。解析过程中，可在词表中查询匹配项。其性能取决于所使用词表的质量和规模。

Arounyadeth Srithirath 等(2013)提出了最长音节匹配和命名实体相结合的混合方法[7]。该方法的基本思想是：首先对文本进行预处理，利用标点符号将输入文本切分成小句，并进行音节提取，然后使用向前算法在词典中查找匹配项，若匹配成功，则将此字段切分出来，若失败，则将这个字段最后一个音节去掉，剩下的字符串再次进行匹配，重复上述过程，直到切分出所有词为止。实验过程中，还加入了命名实体来提升切分效果，编制了老挝语人名和地名构成规则(流程见图 1)。最后，利用从网上获取的新闻语料进行评测，并对是否加入命名实体的两种情况进行对比，结果如表 1 所示。

基于词典或规则的方法简单高效，但具有以下几个缺点：1) 需要语言学家制定相关规则，成本较高；2) 领域移植性较差；3) 对于老挝语这样的低资源(low resource)语言来说，获取高质量的语言资源难度较大；4) 不能很好地解决未登录词和歧义问题，比如“ຂ້ອຍໃຫ້ການສະໜັບສະໜູນເຈົ້າ”(我给你支持)，利用这种方法可能被切分为“ຂ້ອຍ”(我) [代词]、“ໃຫ້ການ”(给予证据) [动词]、“ສະໜັບສະໜູນ”(支持) [动词]、“ເຈົ້າ”(你) [代词]，但是正确的切分应该是“ຂ້ອຍ”(我) [代词]、“ໃຫ້”(给予) [动词]、

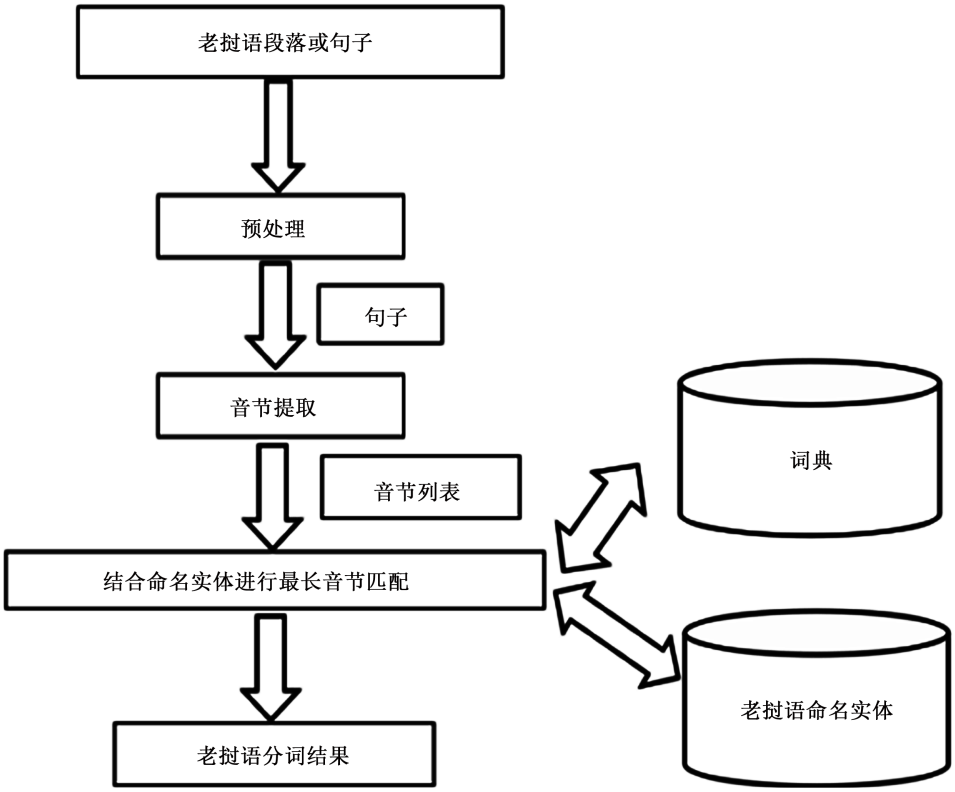

Figure 1. Lao word segmentation system
图 1. 老挝语分词系统

Table 1. The results of the longest syllable matching test before and after the addition of named entity
表 1. 加入命名实体前后最长音节匹配评测结果

方法	准确率(%)	召回率(%)	F-测度值(%)
未加入命名实体	79.87	88.62	84.02
加入命名实体	85.21	92.36	88.64

“ການສະໜັບສະໜູນ”(支持) [名词]、“ເຈົ້າ”(你) [代词]。

B. 基于机器学习的方法

由于基于词典或规则的方法有上述不足，机器学习的方法成为了分词技术的主流。在众多机器学习模型中，CRF 具有特征选取灵活、拟合程度好、训练时间不长等优点。

Sisouvanh Vanthanavong 等(2011)提出了一种基于条件随机场的老挝语分词模型(LaoWS) [8]。该系统的训练数据集包括字母库构成的特征集以及约 10 万词的人工标记语料库(LaoCORPUS) (如图 2)，在该系统中分词问题被当成是字母序列标注任务，词首用字母 B 表示，词中用 I 表示，这样每个字母的标注实际上就是一个二元分类问题。最后，对该系统进行了评测，并与基于词典的方法作了对比(见表 2)。

由对比结果可知，在测试语料相同的情况下，基于条件随机场的方法效果更优，特别是在命名实体的切分上，准确率更高。但是，它依然存在不少问题。其中，最大的问题就是语料库规模不大。而与老挝语类似的泰语在这方面取得的效果远优于老挝语，主要原因就是其采用了 700 万词的语料库，F-测度值高达 96%左右[9]。其次就是切分准确率不高问题。而通过融合词典、规则等方法，可以进一步提高命名实体的切分准确率，并且采用音节而非字母作为特征来训练模型可能也会对改进方法提供帮助。

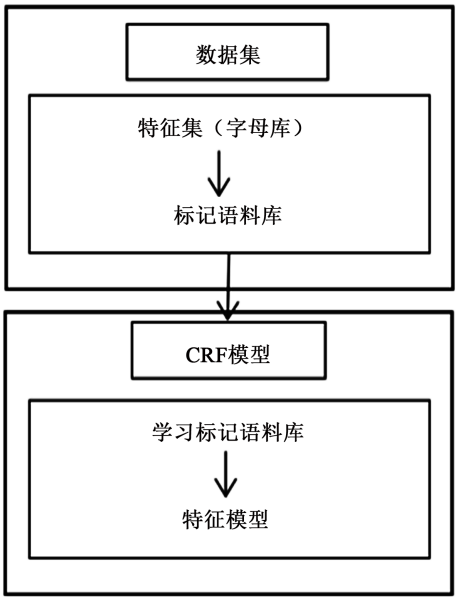

Figure 2. Schematic model of word segmentation
图 2. 分词模型示意图

Table 2. Comparison of participle results between CRF and DCB
表 2. CRF 和 DCB 分词结果对比

方法	准确率(%)	召回率(%)	F-测度值(%)
DCB	80.00	66.67	72.73
CRF	80.29	78.45	79.36

2.2. 老挝语命名实体识别

2.2.1. 命名实体识别和系统性能评测的主要方法

1) 基于规则的方法：该方法主要依赖知识库和词典等资源，以模式和字符串匹配为主要手段。但是这些规则往往依赖于语种、领域和文本风格，制定成本较高且难以涵盖所有语言现象，可移植性不强[10]。

2) 基于统计的方法：统计自然语言处理技术主要来源于机器学习和数据挖掘。该方法使用大型语料库来开发近似语言现象的通用模型。

3) 混合方法：自然语言处理并不完全是一个随机的过程，单独使用基于统计的方法会导致状态搜索空间非常庞大，必须借助规则知识提前过滤。混合方法又可分为规则和统计方法融合、统计方法之间或内部层叠融合等。

系统性能的度量是以数学方式进行的，需要提前准备一个人工标记的测试数据库。主要的衡量指标有准确率(P)、召回率(R)和两者的加权调和平均 F-测度值(F-measure)。

2.2.2. 老挝语命名实体研究成果比较分析

目前，研究老挝语命名实体识别的机构很少[1]，下面将对其主要研究成果进行总结并对比分析。

A. 研究成果一

该成果主要采用了以下三种方法[11]：

1) 基于条件随机场与启发式信息的老挝语人名和地名识别；

[1]据笔者所知，仅昆明理工大学有相关研究成果。

2) 融入广义期望准则[2]的半监督层叠条件随机场的老挝语人名和地名识别；

3) 基于词典与条件随机场的老挝语组织机构名识别。

其中，方法(1)初步实现了简单的老挝语人名和地名识别，具体如图 3、图 4 所示。

这种方法首先利用人工标记的语料库以及人工制定的特征模板训练出命名实体识别模型，然后利用启发式信息进行纠正。

方法(2)完成了复杂、嵌套的老挝语人名和地名的识别：首先采用融入了广义期望准则的单层条件随机场对简单的老挝语人名和地名进行识别并抽取，并在此基础之上，将第一层抽取出的简单老挝语人名和地名作为特征，结合第二层融入了广义期望准则的条件随机场对复杂的、嵌套的老挝语人名和地名再次标注训练，以获取最终的老挝语命名实体识别模型。其中采用了“特征标注”的方法，而不是“样本标注”的方法来训练数据，节省了人工标注的时间(如图 5)。

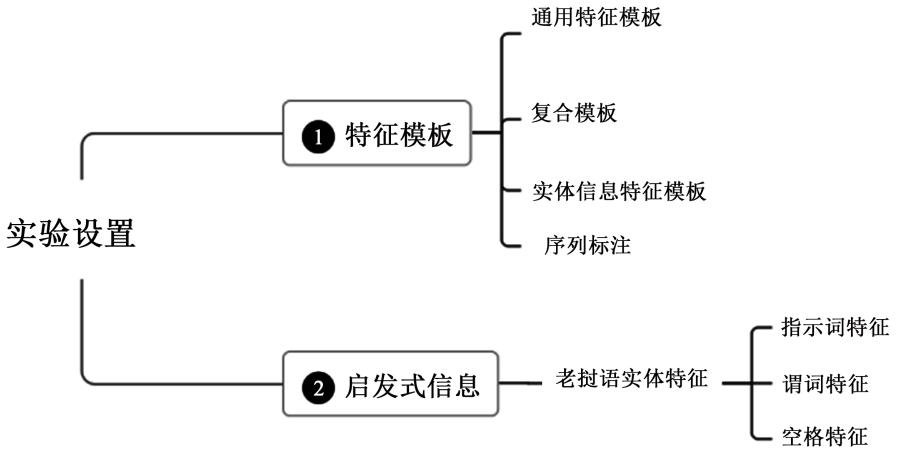

Figure 3. Experimental setup diagram
图 3. 实验设置图

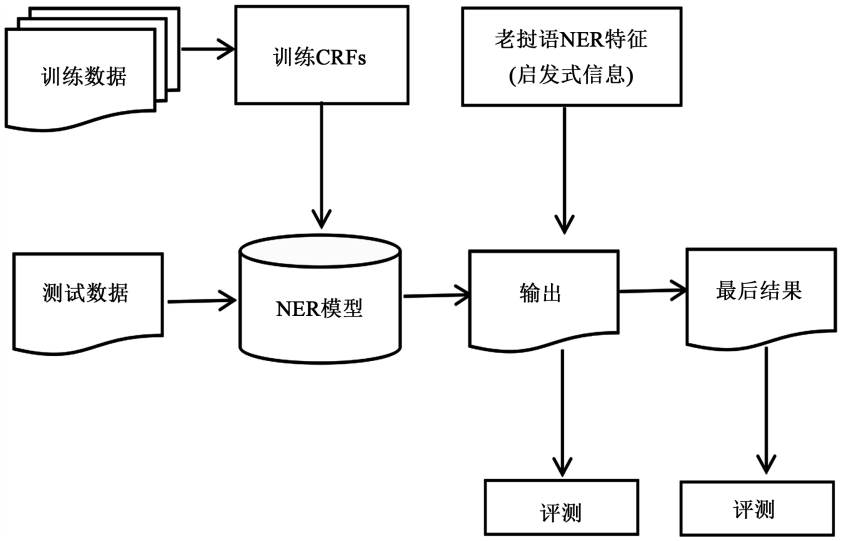

Figure 4. Training flow chart
图 4. 训练流程图

[2]广义期望准则是由 Andrew McCallum 等在 2007 年提出的，它是一个把有关模型期望的优先选择结合到参数估计目标函数中的一个框架中，表达的是模型期望值的优先权。

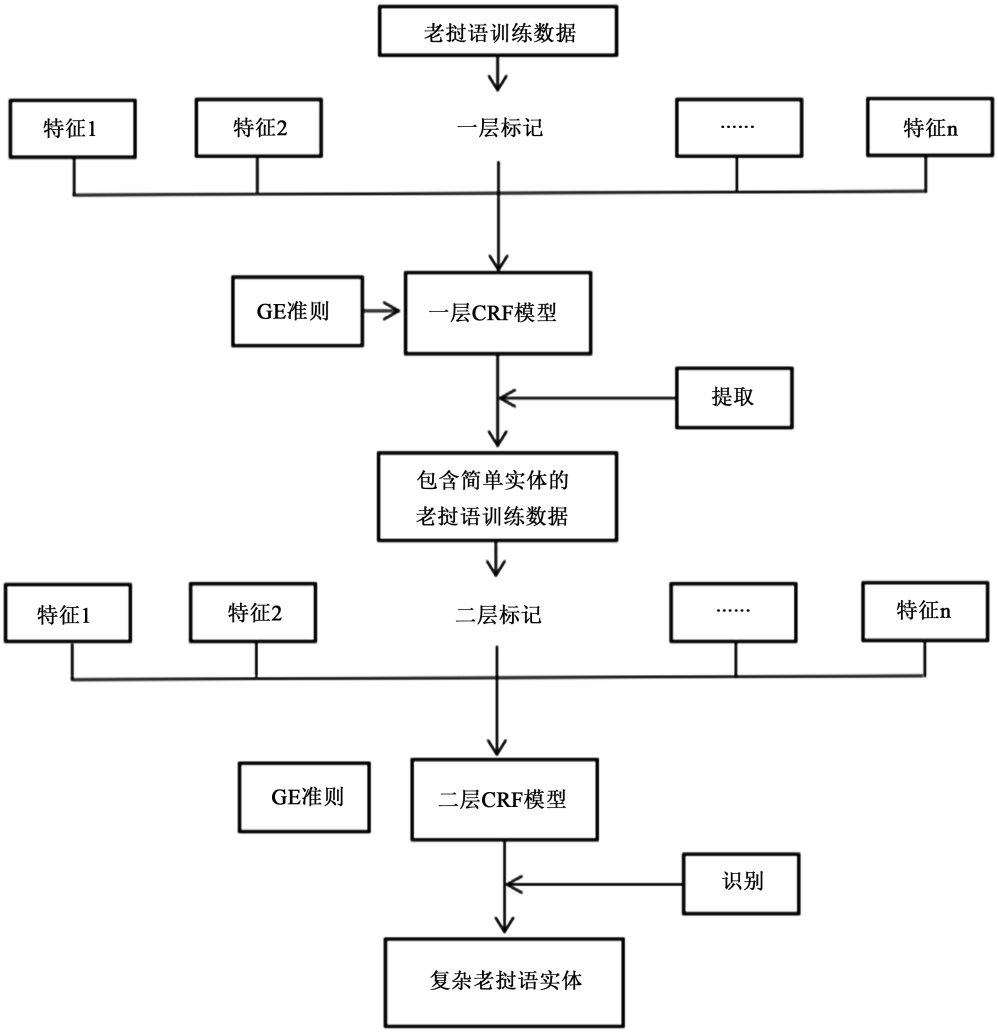

Figure 5. Laminating conditions flowchart random onomastic entity recognition
图 5. 层叠条件随机场人名地名实体识别流程图

方法(3)主要用于组织机构名的识别，这类实体是老挝语中识别难度最大的一种，因为老挝的组织机构名种类繁多，内部构成复杂多变，且存在大量的简称。该方法将老挝语词典及少量老挝组织机构名作为特征融入条件随机场对老挝语组织机构名进行识别(如图 6)。

以上三种方法根据老挝语实际情况和特征，采用混合方法实现了老挝语人名、地名和机构名三类主要命名实体的识别，评测结果汇总如表 3。

由上表可以看出，三种方法在融合其他技术之后，识别效果均有一定程度的提高。其中，融合广义期望准则的半监督层叠条件随机场的方法在识别人名地名方面要优于方法一，因为采用广义期望准则可使用少量的标注语料约束模型对未标注语料进行预测，而有监督的条件随机场需要依赖大量的标注语料才能训练出好的命名实体识别模型，方法(1)虽然采用启发式信息对条件随机场识别出的老挝语候选命名实体进行了纠正，但是候选命名实体有限，而且利用层叠条件随机场对识别复杂的老挝语人名和地名比利用单层的条件随机场效果要好。另外，组织机构名的识别效果与识别人名地名相比还有较大差距。即使加入了词典特征使得组织机构名的各项比率都有了一定程度的提高，但是，其识别效果仍然不是很理想。接下来的研究成果二中对老挝组织机构名的识别作了更为深入地研究，也取得了更好的效果。

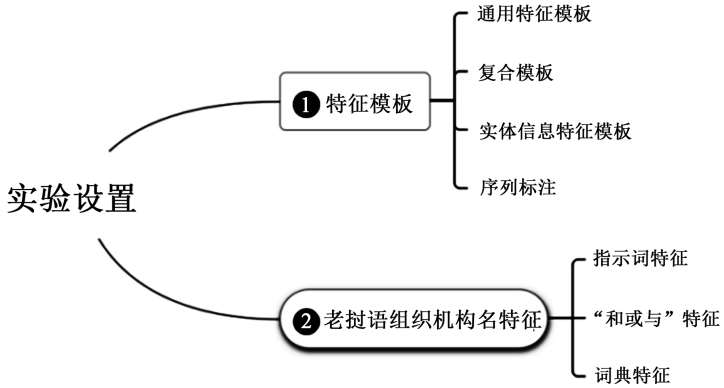

Figure 6. Experimental setup diagram
图 6. 实验设置图

Table 3. Comparison summary table of evaluation results of the three methods[3]
表 3. 三种方法评测结果对比汇总表 [3]

方法	对比	命名实体	准确率(%)	召回率(%)	F-测度值(%)
(1)	CRFs	人名	81.62	83.27	82.44
		地名	72.53	76.24	74.34
	CRFs + 启发式信(选取概率值为 0.6)	人名	83.73	85.23	84.47
		地名	77.61	79.56	78.57
(2)	CRFs (标注实例)	人名	81.62	83.27	82.44
		地名	72.53	76.24	74.34
	层叠 CRFs + GE 准则(标注特征)	人名	85.23	88.01	86.59
		地名	80.54	85.56	83.00
(3)	CRFs	组织机构名	62.30	64.51	63.39
	CRFs + 词典特征		70.73	73.15	71.92

B. 研究成果二

该研究主要采用以下三种方法[12]：

1) 基于分歧的老挝语命名实体识别；

2) 基于层叠条件随机场的老挝机构名识别；

3) 基于条件随机场和支持向量机双层模型的老挝机构名识别。

其中，方法(1)通过基于分歧的方法的对老挝语的人名、地名和组织机构进行了识别：首先通过条件随机场利用实体语料训练 3 个有监督的分类器，然后通过分类器之间的分歧性对未标记语料进行标记。虽然该方法识别效果不如后两种方法，但可以通过该方法获取更多的语料，为后续研究提供支持(如图 7)。

方法(2)构建了一个双层条件随机场，由于有很大一部分老挝语机构名存在嵌套现象，也就是说在老挝组织机构名中可能会包含老挝人名、地名等其他类型的命名实体，例如：“ວິທະຍາຄານປ້ອງກັນຊາດໄກສອນພອນວິຫານ”(老挝凯山·丰威汉国防学院)，其中“ໄກສອນພອນວິຫານ”(凯山·丰威汉)是人名。因此，针对这些复杂的老挝机构名，首先将观察值输入到第一层的条件随机场模型中，再结合第一层模型的识别结果进行第二层识别，这在很大程度上能提高对老挝复杂组织机构名识别的效果(如图 8)。

[3] 方法(1)中“CRFs + 启发式信息”选用概率值为 0.6 时的评测结果，方法(2)选用样本个数为 2000 时的评测结果。

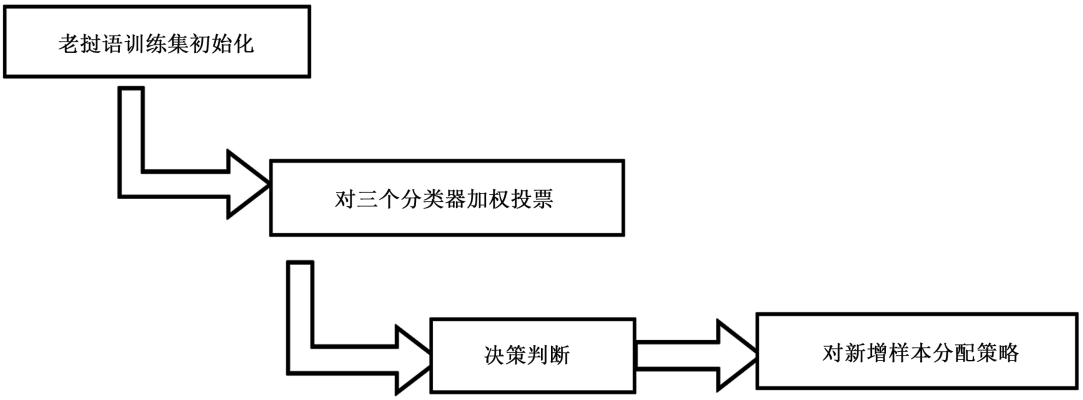

Figure 7. Experimental flow chart
图 7. 实验流程图

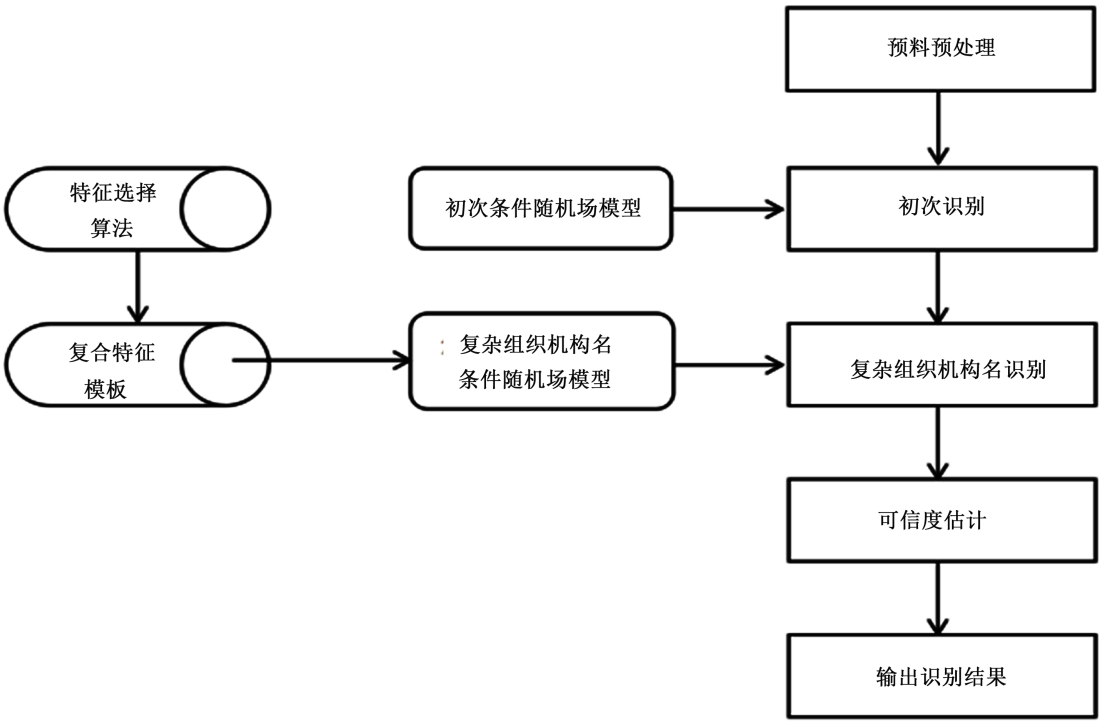

Figure 8. Experimental flow chart
图 8. 实验流程图

方法(3)总结了前两种方法的优缺点，并作了改进。方法(1)在一定程度上解决了老挝语命名实体语料稀缺、获取速度较慢等困难，方法(2)针对老挝机构名提出了一种基于层叠条件随机场模型的识别算法。但是由于制定特征模板时总结的特征有限，识别效果还有待提高。同时，研究人员在深入分析老挝机构名的构成特点后，发现大部分老挝机构名都有一个边界特征词，如果通过专门识别老挝机构名的边界特征词进而识别老挝机构名，准确率会有所提高。方法(3)进行实验前，需要准备老挝语机构名特征词表[4]、后部词表[5]、左右指界词表[6]、简单机构名表以及标记集，实验流程如图 9。

[4]指机构名中具有特殊含义的词，如“工厂(ໂຮງງານ)、大学(ວິທະຍາໄລ)、公司(ບໍລິສັດ)”等。
[5]指除老挝机构名中特征词之外的词，其中老挝地名性名词和普通名词的所占比重很大，结构复杂，随机性很强。
[6]左指界词指出现在老挝机构名前面的第一个词，右指界词是指出现在老挝机构名后面的第一个词。

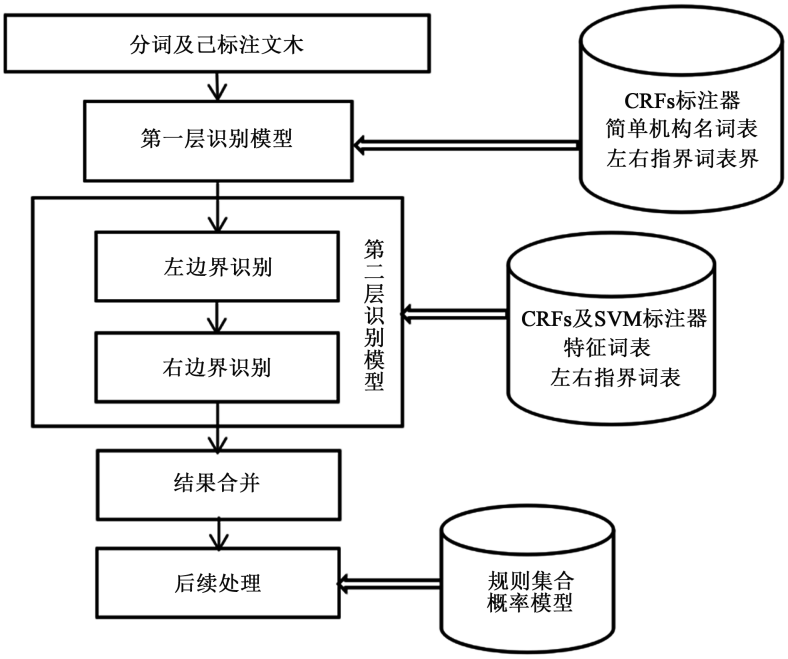

Figure 9. Experimental flow chart
图 9. 实验流程图

老挝语机构名的识别归根到底是一个二元分类问题：把特征词表中出现的词都标记为左边界的候选词。然后，应用支持向量机模型对候选边界词进行筛选，确认其是否为老挝机构名的左边界词。左边界确定后，用 CRF 模型进行后部标注。在选用标注策略的时候，考虑到使用全标策略可能会造成大量的浪费，因此方法(3)使用基于驱动的标注。其中，后续处理分为两部分，第一部分构建概率模型，进行可信度计算，主要对识别结果中低于设定阈值的字符串进行计算。假如可信度高于此阈值，则标记为老挝语机构名。第二部分则利用老挝语特征构建的规则模型对部分不完整的老挝机构名进行识别。

以上三种方法的研究实际上是一个既循序渐进又相辅相成的过程，尤其针对老挝组织机构名的识别使用混合方法作了研究，取得了较好的效果。三种方法的评测结果汇总如下表 4。

C. 命名实体研究成果小结

由上述分析可知，目前老挝语命名实体识别主要采用以基于统计为主的混合方法，因为单独使用基于统计的方法会使状态搜索空间非常庞大，影响识别效率，所以必须借助规则提前进行过滤修剪处理。而其中采用的统计方法又以条件随机场居多，主要原因是该方法简单易行，而且效果较好，能够为命名实体识别提供一个特征灵活、全局最优的标注框架，可以说是命名实体识别中最成功的方法。但是，同英语和汉语等语种相比，老挝语命名实体识别效果仍然还有很大的提升空间，所面临的难点既有大部分语种都会遇到的共性问题，也有其个性问题。具体如下：

1) 老挝语文本没有天然的词边界和明显的形态标志，而英语单词之间就有空格并且实体名称的首字母一般大写；

2) 命名实体识别与老挝语分词、浅层语法分析等过程相互影响、相互制约；

3) 老挝语语言资源匮乏，没有大型语料库和专名库等基础资源；

4) 随着经济社会的发展，专有名词范围扩大，不仅限于人名、地名和机构名，电影名、书名、项目名以及高新技术领域名词等也包含在内；

5) 老挝语层次划分颗粒度越来越小，比如地名被细化为国家、省、市、县、村等；

Table 4. Comparison summary table of evaluation results of the three methods
表 4. 三种方法评测结果对比表

序号	方法	命名实体	准确率(%)	召回率(%)	F-测度值(%)
(1)	基于分歧	人名	73.44	75.65	74.53
		地名	72.53	74.24	73.39
		机构名	71.23	73.25	72.24
(2)	基于层叠 CRF 模型 [7]	机构名	77.72	79.67	78.68
(3)	基于 CRF、SVM 双层模型	机构名	80.83	82.75	81.75
			70.73	73.15	71.92

6) 新的命名实体不断涌现，但语料老旧，覆盖不全，有大量新兴词汇和外来词，并且对外来词的翻译没有统一的规范；

7) 老挝语命名实体中常出现老挝语英语交叉使用等情况，这使得老挝语命名实体识别的任务大大增加；

8) 老挝语命名实体歧义现象众多，消歧困难；

9) 老挝语中存在大量缩略词，且经常出现一对多或多对一的情况。

3. 老挝语命名实体识别的发展趋势

以上研究成果基本都采用了有监督的机器学习方法，如 CRF、SVM 等，但是由于老挝语标注语料稀缺，小规模的数据很难实现泛化的识别效果。虽然针对不同的实体类别刻画了大量不同的人工特征，如词边界特征、构词特征等，这些特征具有一定的针对性，能够较为准确地反映识别对象的特点，在一定程度上弥补老挝语语料稀缺的问题，但是人工特征耗时耗力，成本较高，并且不同识别对象特征不尽相同，可移植性不强。

近年来，基于神经网络的深度学习技术发展迅猛，为语音识别、图像识别、自然语言处理等提供了强大的工具，也为这些领域的发展提供了新的契机[13]。英语、汉语等语种已经将这项技术应用到了命名实体识别中，并取得了很好的效果。这种方法的核心技术是词向量，词向量是通过训练神经网络语言模型得到的一种分布表示特征。这项技术将词转化成为稠密向量 [8]，并且相似的词对应的词向量相近。以词向量为核心的深度学习模型比传统的机器学习模型有更强的特征抽取能力，并且抽取出的特征包含的信息量也更加丰富，可以进一步减少对人的依赖[14] [15]。如果将这一技术引入到老挝语命名实体识别当中，替代传统的机器学习模型，应该会取得更好的效果。

命名实体识别就是一个序列标注问题，而这个问题利用循环神经网络(Recurrent Neural Network, RNN)就可以很好地解决。但是，RNN 也存在着不足，其中之一就是它无法很好地处理长距离依赖问题。但是，利用长短时记忆(Long Short-Term Memory, LSTM)模型就可以解决这个问题[16] [17]。总之，利用基于深度学习的神经网络可以减小人工构建特征的所付出的代价，提高特征抽取效率，利用尽可能少的语料挖掘尽可能多的数据信息，进而进一步提升老挝语命名实体识别效果。

但是，目前该方法仍存在一定的问题：

1) 不能准确描述词向量包含了哪些信息(如每维特征代表的意义)；

2) 词向量的训练采用无监督方式，不能很好的利用先验信息。而如果在该方法的基础上加入老挝语

[7] 取六次实验平均值。
[8] 即词汇的稠密向量化表示(dense word embedding)。

实体特征和领域知识，则可能会取得更大的突破；

3) 词向量是神经网络语言模型的副产物，其损失函数不是由具体应用构建的。因此，不是词向量训练的越好，应用效果就越好。

此外，随着老挝网络技术的发展，将会有海量的老挝语数据产生，如何在互联网上利用自然语言处理技术捕捉重要信息将会成为今后研究的重点。除了文本，还会有大量的图片、音频和视频等多模态数据。因此，将老挝语自然语言处理技术与语音识别、图像处理等领域知识相结合也将成为老挝语命名实体识别的必然趋势。

4. 结语

命名实体识别是自然语言处理中最基本也是最重要的任务之一，直接影响着信息检索、信息抽取、句法分析、机器翻译等领域的研究和应用。命名实体识别的方法可以归纳为三类：基于规则、基于统计以及基于混合的方法。目前，大多数命名实体识别技术都是基于混合方法的——以统计方法为主，加入人工编制的规则，并达到了不错的效果。老挝语命名实体识别的已有成果也是沿用这种方法。但是要想进一步提高识别效果，还有一系列的困难和挑战：老挝语单词之间无空格、首字母不大写等先天处理劣势，以及语料稀缺、领域专家少等后天薄弱环节。除了加大语料库建设力度等措施以外，势必要引入更为先进的技术，近年来流行的基于深度学习的神经网络方法将是未来研究的重点。

基金项目

资助课题：1) 国家自然科学基金项目(U1204602)；2) 数学工程与先进计算国家重点实验室开放课题项目(2013A14)。

参考文献

[1] Rau, L.F. (1991) Extracting Company Names from Text. *Seventh IEEE Conference on Artificial Intelligence Applications*, Miami Beach, 29-32.

[2] Sundheim, B., *et al.* (1996) Message Understanding Conference-6: A Brief History. *Conference on Computational Linguistics*, Association for Computational Linguistics, Philadelphia, 466-471.

[3] Nadeau, D. and Sekine, S. (2007) A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes*, **30**, 3-26.

[4] 银莎格. 国内老挝语研究综述[J]. 铜仁学院学报, 2014(1): 113-116.

[5] 宗成庆. 统计自然语言处理[M]. 北京: 清华大学出版社, 2013: 152.

[6] 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.

[7] Srithirath, A. and Seresangtakul, P. (2013) A Hybrid Approach to Lao Word Segmentation Using Longest Syllable Level Matching with Named Entities Recognition. *International Conference on Electrical Engineering/Electronics*, *Computer*, *Telecommunications and Information Technology*, Krabi, 1-5.

[8] Vanthanavong, S., Haruechaiyasak, C. and Lao, W.S. (2011) Lao Word Segmentation Based on Conditional Random Fields.

[9] Haruechaiyasak, C., Kongyoung, S. and Dailey, M. (2008) A Comparative Study on Thai Word Segmentation Approaches. *International Conference on Electrical Engineering/Electronics*, *Computer*, *Telecommunications and Information Technology*, *Ecti-Con*, Krabi, 125-128.

[10] 孙镇, 王惠临. 命名实体识别研究进展综述[J]. 数据分析与知识发现, 2010, 26(6): 42-47.

[11] 杨梦杰. 老挝语命名实体识别方法的研究[D]: [硕士学位论文]. 昆明: 昆明理工大学, 2016.

[12] 段韶鹏. 老挝语命名实体识别研究[D]: [硕士学位论文]. 昆明: 昆明理工大学, 2017.

[13] 奚雪峰, 周国栋. 面向自然语言处理的深度学习研究[J]. 自动化学报, 2016, 42(10): 1445-1465.

[14] 麻泽武. 基于深度学习的命名实体识别技术研究[D]: [硕士学位论文]. 杭州: 浙江大学, 2016.

[15] 张海楠, 伍大勇, 刘悦, 等. 基于深度神经网络的中文命名实体识别[J]. 中文信息学报, 2017, 31(4): 28-35.

[16] 冯艳红, 于红, 孙庚, 等. 基于 BLSTM 的命名实体识别方法[J]. 计算机科学, 2018, 45(2): 261-268.

[17] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735