doi:10.3969/j.issn.1673-9833.2020.03.004

菲律宾语自然语言处理研究综述

李珊珊[1,2]，蒋盛益[1,2]，符斯慧[2]

（1. 广东外语外贸大学 广州市非通用语种智能处理重点实验室，广东 广州 510006；
2. 广东外语外贸大学 信息科学与技术学院，广东 广州 510006）

摘 要：通过对菲律宾语的词法分析、句法分析、语义分析等基础研究和机器翻译、拼写检查、情感分析等应用技术的研究进展进行分析，得知菲律宾语仍属于语言资源较为缺乏的低资源语言，在菲律宾语自然语言处理领域，现有研究比较宽泛但不深入，与英语、汉语等语种的自然语言处理研究相比，还存在较大差距；相较而言，英菲平行语料库构建及其机器翻译的研究取得了较大进展，而其他领域研究进展相对缓慢。总体来说，通过跨语言处理技术构建跨语言平行语料库，推动深度学习应用于菲律宾语自然语言处理的方法研究，探讨基于规则、图模型、结构等方法对菲律宾语文本自动摘要的适用性，将是未来菲律宾语自然语言处理的主要研究方向。

关键词：菲律宾语；黏着语；低资源语言；自然语言处理；词性标注

中图分类号：TP312 **文献标志码**：A **文章编号**：1673-9833(2020)03-0023-10

引文格式：李珊珊，蒋盛益，符斯慧 . 菲律宾语自然语言处理研究综述 [J]. 湖南工业大学学报，2020，34(3)：23-32.

An Overview of Natural Language Processing of Filipino

LI Shanshan[1,2]，JIANG Shengyi[1,2]，FU Sihui[2]

（1. Guangzhou Key Laboratory of Multilingual Intelligent Processing，Guangdong University of Foreign Studies，Guangzhou 510006，China；2. School of Information Science and Technology，Guangdong University of Foreign Studies，Guangzhou 510006，China）

Abstract：Based on the analysis of morphological analysis, syntactic parsing, semantic analysis, along with the research progress of machine translation, spelling check, sentiment analysis and other application technologies, a conclusion can be drawn from it that Filipino is still a language with a relatively lack of language resources. As regards to the natural language processing of Filipino, the existing research is relatively broad but not in-depth, compared with that of English, Chinese and other languages, with a big gap in the research of natural language processing. Compared with other research areas, great progress has been made in the construction of parallel corpus and machine translation for English-Filipino, with a relatively slow progress in other fields. In general, the main research direction of NLP for non-native Filipino is for the construction of a parallel corpus based on cross language processing technology, the promotion of the research of deep learning applied to NLP, and the exploration of the applicability of rules, graph

收稿日期：2020-04-02

基金项目：国家自然科学基金资助项目（61572145）

作者简介：李珊珊（1995-），女，广东遂溪人，广东外语外贸大学硕士生，主要研究方向为数据挖掘和自然语言处理，E-mail：2209222731@qq.com

通信作者：蒋盛益（1963-），男，湖南隆回人，广东外语外贸大学教授，博士，主要从事数据挖掘和自然语言处理方面的教学与研究，E-mail：jiangshengyi@163.com

models, structures and other methods to the automatic abstracts of Philippine texts.

Keywords：Filipino；agglutinative language；low resource language；natural language processing；part-of speech tagging

1 研究背景

作为菲律宾官方语言的菲律宾语，又称他加禄语，属于南岛语系的马来－波利尼西亚语族，主要被使用于菲律宾，也广泛运用于马来西亚沙巴州、印度尼西亚北部地区和新加坡。在菲律宾人口中，超过 2 000 万人以菲律宾语作为母语。菲律宾语采用的书写系统为拉丁字母。在菲律宾语的发展过程中，因受殖民统治和外来文化的影响，其从西班牙语、福建闽南话、英语、马来语、阿拉伯语等语言中吸收了不少词汇。菲律宾语属于黏着语（agglutinative language），但是也呈现出一些屈折语的特征，如动词的形态受焦点、体及语态的影响，代词的形态受数的影响等。菲律宾语的词汇形态变化复杂，句法结构复杂，且单词顺序较为自由。

菲律宾是东南亚的一个发展中国家，也是“一带一路”沿线的重要国家之一。1975 年中菲建交以来，两国关系总体上发展顺利，各领域的合作不断被拓展。随着“一带一路”倡议提出之后，中菲两国在政治、经济贸易、文化等领域的合作有了进一步的发展，致力于共同深化和平与发展的战略性合作关系。中菲两国在文化交流与合作更加密切的同时，语言互通的需求也日渐强烈。在当今互联网快速发展的时代，如何利用信息技术，构建“语言互通”的桥梁，进一步深化我国与菲律宾国家的文化与信息交流，促进区域合作，实现共同发展，显得十分必要。为此，有不少学术研究团队以菲律宾语为对象进行学术研究，主要的研究团队包括菲律宾德拉萨大学语言技术中心（De La Salle University，Center for Language Technologies）、菲律宾理工大学计算机与信息科学学院（Polytechnic University of the Philippines，College of Computer and Information Sciences）、广州市非通用语种智能处理重点实验室（Guangzhou Key Laboratory of Multilingual Intelligent Processing） 等。由此可见，对菲律宾语自然语言处理方面展开研究具有重要的现实意义。因此，本文拟对菲律宾语的词法分析、句法分析、语义分析等基础研究和机器翻译、拼写检查、情感分析等应用技术的研究现状进行归纳与分析，并且梳理已有资源建设的研究成果，剖析其面临的主要问题，在此基础上展望其未来的研究方向。

2 菲律宾语自然语言处理现状

2.1 词法分析

词法分析研究主要包括词干提取、形态分析（如词形还原等）、词性标注等基础研究，以及命名实体识别等应用技术。本小节介绍的内容仅涉及菲律宾语自然语言处理领域的底层技术，如词干提取、形态分析、词性标注等，而其他应用技术研究内容将在后续章节中展示。

2.1.1 形态分析

菲律宾语的词缀系统非常复杂，包含前缀、中缀、环缀、后缀、重复及以上多种词缀的叠加。菲律宾语中的重复可以是单词部分重复或者全部重复。多种词缀叠加是菲律宾语动词中常见的语言现象。例如单词 pinanglilibang-libang，它通过词干 libang 附加前缀 pang，而前缀 pang 又叠加中缀 in 组成 pinang，并且部分重复 li 及全部重复 libang 来构成。由于菲律宾语的复杂性，对其进行形态分析成为菲律宾语自然语言处理领域的基础任务，可为信息检索、机器翻译等研究提供支持。

F. Fortes[1] 提出一个用于提取级联（concatenative，指包含前缀、后缀、环缀的情况）和非级联（non-concatenative，指包含中缀、重复的情况）形式的词干动词词法分析器——TagMA。TagMA 通过语素、CV（C，consonants 表辅音；V，vowels 表元音）以及音节来表示输入的动词，再将输入表示馈入生成器以得到候选集合。虽然 TagMA 的分析准确率达 96%，但是利用该方法分析输入动词的过程耗时较长，且只输出动词的词干、词缀和时态，而不包括动词的不定形式。

在此基础上，F. C. Fortes-Galvan 等 [2] 将最优化理论（optimality theory）应用于词法分析中，提出了一个基于约束的动词词法分析器，以同时处理级联和非级联的形态学现象。并且课题组利用该分析器对含有 50 个词根的 1 600 个动词进行了测试，所得结果表明，分析器所输出的动词基本形式的准确率达 96%。

R. Roxas 等 [3] 也设计了一个动词词法分析器。与 F. C. Fortes-Galvan 等提出的分析器不同的是，R. Roxas 设计的分析器输出的结果中包含动词的时态、不定式

形式以及词缀。对于一个给定的动词，该词法分析器能给出该动词的基本形式、所含词缀以及对应的时态（过去时、现在时及将来时）。利用该分析器对 1 050 个动词（包含规则动词及不规则动词）进行测试，所得测试结果表明，其对于 3 种输出结果的准确率均达 95% 以上。

以上两项研究只是对动词进行了分析和还原，而 D. E. Bonus[4] 提出了一个不限于动词的基于词典的词根还原算法 TagSA（https://github.com/laronandrew11/stemmer），该算法中考虑了词缀、重复以及复合等情况。并在 6 000 多个词语上进行了测试，且取得了不错的效果。

P. Baumann 等 [5] 研究了如何利用语言资源丰富的英语来辅助对于资源缺乏的菲律宾语及祖鲁语的形态进行分析。他们考虑到这两种语言的形态变化较为丰富，并且由于受到外来文化的影响，有不少借词现象，因此，可以根据借词的形态变化来获取常用的词缀。以获取菲律宾语单词的词缀为例，他们先从网上获取两种语言的文本，并且分别从中提取出对应的词汇列表；再通过判断某个英语单词是否为一个菲律宾语单词的子串，以获得潜在的词缀。最后，根据潜在的词缀在语料中的分布，确定最终的词缀。他们利用该方法成功提取出 28 个常用的菲律宾语词缀以及 66 个常用的祖鲁语词缀。

通过以上的研究分析可以看出，由于菲律宾语的动词形态变化较其他词类的更为丰富，其形态分析研究主要针对动词，较少有研究针对所有词类。几乎所有的研究是利用菲律宾语动词的形态学变化规律提出基于规则的形态分析方法，算法的准确率也较高。本课题组认为，虽然菲律宾语动词的形态变化复杂，但是均有规律可循，可以通过构建大规模（词干、派生词）序列对语料库，将形态分析任务转化为序列学习任务，通过深度学习方法，如 LSTM（long short-term memory）、seq2seq 等，可自动学习菲律宾语动词的形态学规则，从而实现词干的提取。

2.1.2　词性标注

词性（part-of-speech，POS）是词汇最基本的语法属性，使用词性标注便于判定每个词的语法范畴。词性标注是自然语言处理中一项非常重要的基础性工作，其为句法分析、命名实体识别、机器翻译等任务打下基础。与英语相比，菲律宾语同样具有后缀、大写字母等可用于确定 POS 的语言特征。除此以外，菲律宾语的词性标注离不开前缀、中缀、环缀、重复等有用的语言信息。

Cheng C. K. 等 [6] 提出了一个基于模板的 *n* 元语法词性标注器，其核心为几类词特征，即常用的 225 个用于构建句子的词语、词缀、字母大写以及连字符。他们所用的训练和测试语料源于菲律宾语版圣经（共 141 句），用到的词类标签有 59 个，测试结果的准确率为 92% 以上。

M. Erlyn 等 [7] 探讨了影响菲律宾语词性标注效果的因素，考虑以菲律宾语单词的形态结构、形态信息（如词缀）作为训练 POS 模型的输入。实验中使用了菲律宾德拉萨大学（De La Salle University，DLSU）的人工标注数据，涵盖小说、报纸文章、短片故事和圣经章节，包括 114 096 个词条，POS 标注集包括 9 个粗粒度标签、60 个特定标签、5 个标点符号标签以及其他符号的标签，所得测试结果表明，POS 模型标注的准确率高达 93% 以上。

C. D. E. Reyes 等 [8] 利用支持向量机和 bigram 开发了一个菲律宾语词性标注器 SVPOST，并对其有效性进行了实验验证。其实验数据中包含 122 318 个已标注单词和 64 个词性标签。实验结果表明，该标注器的准确率可达 81%。

N. Nocon 等 [9] 将统计机器翻译的方法应用于菲律宾语的词性标注中。他们将序列标注问题转换为编码－解码问题，并以给定的句子（源语言）作为输入，句子中的词语对应的词性标记（目标语言）为模型的输出。实验中使用的词类标记集为 MGNN 标记集（包含 230 个词类标记，http://goo.gl/dY0qFe），所用的训练和测试语料取自维基百科（共 2 668 句），得到的最高准确率为 84.75%。

M. P. Go 等 [10] 构建了基于 Stanford 词性标注器的菲律宾语词性标注（https://github.com/matthewgo/FilipinoStanfordPOSTagger）。他们用到的核心算法为最大熵循环依赖网络，在设计特征时考虑了词汇的形态及句子内部的语码转换信息，使用的词类标记集也是 MGNN 标记集，所用的训练和测试语料来源于英文维基百科随机抽取的 15 166 个句子，经由相关语言学家翻译为菲律宾语句子后再进行人工词性标注，最终得到的标记准确率为 96%。

J. F. T. Olivo 等 [11] 尝试了基于条件随机场的方法，使用的词类标记集仍为 MGNN 标记集，所用训练和测试语料与 M. P. Go 等 [10] 所用的语料一致，得到的标记准确率在 90% 以上。

菲律宾语句子中单词顺序自由，导致菲律宾语不可以通过分析目标词前后词汇的分布概率来预测目标词的 POS 标签，将 POS 标注视为序列学习任务则无法很好地学到菲律宾语语法结构模式，从而导致实验效果不好；而标注语料的缺乏也限制了词性标注工

作的开展。

2.2 句法分析

句法分析的主要任务是为了确定句子中各组成成分之间的关系，也就是确定其句法结构。菲律宾语的句子中，各组成成分的顺序较为自由，不具有主谓一致的语法特点，并且句子的焦点成为主题而不是主语。这些语言特征成为菲律宾语句法分析中的一大障碍，导致适用于菲律宾语句法分析的算法相对较少，其研究成果也很少。

A. Clark[12] 尝试了利用词汇功能语法（lexical functional grammar，LFG）作为计算模型来捕获菲律宾语的信息，实现了一个用于菲律宾语书面句子语法分析并输出句子功能结构的系统——FiSSAn。虽然 FiSSAn 目前只能用于处理陈述句，但是可以通过总结更广泛的语法规则集以捕获更多类型的菲律宾语句子结构，如祈使句和疑问句等。

D. L. Alcantara 等 [13] 使用无监督的统计方法，对菲律宾语句子进行了构成成分（constituent）的划分。他们在对句子进行词形还原和词性标注后，统计分析所有出现的词性标注序列，以生成划分构成成分的规则，由此得到的规则库即可以用于划分后续句子的构成成分，此方法的 *F* 值在 69% 以上。

E. Manguilimotan 等 [14] 首先进行了针对菲律宾语依存句法分析的研究。他们采用基于图的最大生成树算法，探索了粗细粒度的词性、词根和形态等特征对句法分析模型性能的影响。并且在 2 741 个句子上进行了训练和测试，结果表明，对于无标签的依存关系（unlabeled attachment scores，UAS），句法分析模型的平均准确率为 78%；而对于整个句子，句法分析模型的平均准确率仅为 24%。这一实验结果表明，当词性信息不够准确时，加入形态信息有利于提高句法分析器的性能。

2.3 语义分析

对于不同的语言单位，语义分析有着不同的意义。在词汇的层面上，语义分析指词义消歧；在句子的层面上，语义分析指语义角色标注；在篇章的层面上，语义分析指共指消解。语义分析是目前 NLP（natural language processing）研究的一个重要方向。部分学者对于菲律宾语语义分析进行了初步的探讨和研究，这些研究主要集中在语义知识库的构建、词义消歧等方面。

E. Domingo 等 [15] 研究了将句法关系信息融合到机器翻译系统中，以进行目标语言的词义消歧。他们一方面利用双语词典和 WordNet 进行源语言的词义消歧，另一方面从目标语言词典和语料中统计抽取出句法的关系信息，两者结合以在生成目标语言时选择出最合适的词语。

M. Mistica 等 [16] 初步实现了基于条件随机场（conditional random field，CRF）的语义分析器，以识别菲律宾语中的谓词－论元结构。他们构建了一个小规模的谓词－论元菲律宾语语料库，并且在实验过程中对比了词性、词语形态及字母 *n*-gram 等特征对分析器性能的影响。实验结果表明，对于谓词的识别，*F* 值最高为 44.2%，而对于论元的识别和依附，*F* 值最高为 47.7%。

S. Bergsma 等 [17] 针对附加前缀的动词，提出如果前缀动词可以被分解为包含其词干的语义等效表达，则可认为该词是组成动词。他们还开发了一个分类器，以通过一系列词汇和其分布特征来预测词汇的组成。实验结果表明，该分类器可以较为准确地预测附加前缀的动词的词干。

A. L. Andrei[18] 试图构建了一个小规模的面向 Twitter 的菲律宾语情感词典 LIWC（linguistic inquiry and word count）。首先，他在菲律宾国内的博客、新闻网站及 Twitter 上获取菲律宾语文本，并且通过文本预处理得到了 18 254 个词，其中包含英语、菲律宾语、宿雾语、印尼语和西班牙语等语言的单词。然后，其利用谷歌翻译，将所有词翻译为菲律宾语词，经过人工校对过滤后，得到了 1 510 个菲律宾语词；再仿照构建英语 LIWC 的步骤，让 3 位标注员对所有词进行正负向情感标注，最终获得 273 个正向情感词及 344 个负向情感词。另外，人工标注筛选了大量针对某个话题的推文，最终获得 575 篇带有情感标记 (正向、负向及中性) 的推文，基于这些推文测试了情感词典的效果，在正向文本上的平均 *F* 值为 33%，在负向文本上的平均 *F* 值为 52%，而在中性文本上的平均 *F* 值为 12.5%，说明仍有较大的提升空间。

综上所述，相比词法分析及句法分析等方面的研究，菲律宾语语义分析的研究成果较少，而且其语义知识库构建仍处于初级阶段。

2.4 机器翻译

菲律宾的机器翻译始于 20 世纪 90 年代后期，涉及菲律宾国家的两种官方语言：菲律宾语和英语。截至目前，菲律宾语的机器翻译研究取得了较大进展，其研究方法涵盖基于转换、基于语料库、基于统计和基于深度学习的方法。

最早被用于菲律宾语机器翻译研究的方法是转换法，该方法主要是通过对源语言进行分析，得到其结构，再将分析的结构转换成目标语言的结构，而后

根据目标语言结构生成目标语言，从而实现翻译。例如 R. Roxas 等 [19] 利用增强过滤网络和少于 10 000 词条的字典构建了英菲翻译工具，但该工具仅是针对陈述句和祈使句的翻译。随后，A. Borra[20] 探讨了将词汇功能语法作为文法形式的翻译系统，发现功能结构（f-structure，f 结构）和组分结构（c-structure，c 结构）有助于识别翻译错误。在此基础上，A. Borra 等 [21] 也提出了一个基于词汇功能语法的英菲机器翻译系统。整个系统包括对源语言 f 结构的分析、源语言的 f 结构到目标语言的 f 结构的转换，以及由目标语言的 f 结构生成目标语言几个步骤。在系统开发过程中，用到了两种语言的语法规则、单语词典、转换词典（包含 2 000 个平行词对）及转换规则等语言资源。实验结果表明，输入和输出的句子符合既定的语法规则、其单词存在于词典中且转换规则必须存在相应的 f 结构才可以成功翻译。T. Allman 等 [22] 开发了一个称为 Linguist’s Assistant 的自然语言生成器，可被用于翻译宗教文本。其虽然需要复杂的短语结构规则才能正确地对目标语言的成分进行排序，但是短语生成规则明显简化了目标语言的语法规则。以上基于转换的方法中，翻译的效果受限于语料规模及转换规则，无法翻译词典外的词汇（out of vocabulary，OOV）。

鉴于基于转换方法的人工构造规则的局限性，基于语料库的机器翻译方法应运而生。该方法和传统的基于规则的方法相比有很大的不同，基于语料库的方法并不对目标语言进行深入复杂的语法分析，也不通过规则转换，而使用源语言和目标语言相对照的双语或多语语料库直接或间接地进行翻译。例如 R. E. O. Roxas 等 [23−24] 提出了基于转换规则和基于语料库混合的方法。其中，利用 LFG 实现基于转换的方法，而基于语料库的方法尝试从大量英菲平行句对（包含 207 000 菲律宾语词汇）中抽取翻译模式，并且存为模板，以实现翻译。E. Ong 等 [25] 提出一种基于模板的机器翻译系统，该系统从给定的双语语料库中提取模板，并以常见的词汇过滤及组块对齐算法来提高提取模板的质量。

基于统计的机器翻译方法是一种间接地使用语料库的机器翻译方法，它是通过双语句对的对齐，分析词汇共现的可能性来计算源语言的某一个词映射到目标语言的一个或多个（或零个）词的概率。例如 J. Ang 等 [26] 构建了一个基于 Moses（http://www.statmt.org/moses/）菲英统计翻译系统——FEBSMT，所用的实验数据来源于 22 031 句旅游领域的英菲平行句对。该系统可以接受用户反馈，并且周期性地汇总反馈数据，以对系统做增量式训练，提升系统性能。

由于自动构建平行语料库方法的可用性，基于深度学习的菲律宾语机器翻译研究取得了一定的进展。A. J. Tacorda 等 [27] 利用 100 000 个英菲平行句对训练 RNN 模型，并集成字节对编码（byte pair encoding，BPE）以减少 OOV 翻译错误。BPE 将一个词条分解成可识别的字符序列。因此，如果已经通过 BPE 识别出训练数据的词干和词缀，则可以识别训练数据中不存在的词条。但是 BPE 无法处理误将词干的字符序列识别为词缀的情况。而针对 OOV 翻译的问题，A. N. Lazaro 等 [28] 提出通过利用领域适应技术预处理训练数据，从而减少 OOV 的概率。

菲律宾语除了具有句子结构成分顺序自由的特点外，其动词拥有时态和焦点的特点及词缀包含前缀、中缀、后缀、环缀及重复等复杂的形态变化特点，这些都给菲律宾语机器翻译带来一定的挑战。由于菲律宾语目前还没有成熟可用的语言工具，如词干提取、词性标注等工具，故菲律宾语机器翻译仍有很大的探索和研究空间。

2.5　情感分析

随着互联网技术的普及，越来越多的用户在互联网（如 Twitter、Facebook 等）上发表对于诸如人物、事件、产品等有价值的评论信息。为了理解和分析可能包含用户情感、观点和信念的大量数据，情感分析工作显得至关重要。

R. V. J. Regalado 等 [29] 研究了菲律宾语文本的主观性分类。他们以 TF−IDF 为主要特征，分别对文档和句子用 C4.5、朴素贝叶斯、KNN（k-nearest neighbor）和 SVM（support vector machine）等算法进行了主观性分类。对于文档级别，给出算法中 SVM 算法取得了最高的准确率，为 95.06%；而对于句子级别，朴素贝叶斯算法取得了最高的准确率，为 58.75%。M. Pippin 等 [30] 尝试对菲律宾人发的推文进行了情感分类。他们的情感分类体系中包含 7 个类别：开心、伤心、愤怒、惊恐、惊奇、厌恶及中性。他们用朴素贝叶斯算法在 300 000 篇推文（其中“中性”占最大比例，为 79%；“开心”第二，占 18%）上进行测试，分类准确率约为 70%。

F. Patacsil 等 [31] 获取了菲律宾国内一些热门博客的评论，以研究菲律宾国民对国内 3 家主要因特网服务提供商（intenment server provider，ISP）的看法。他们以 *n*-gram 模型作为主要特征，辅以一些规则，对比了朴素贝叶斯和 SVM 的性能。实验结果表明，使用二元模型的 SVM 获得的情感分析效果较好。

F. R. Lapitan 等 [32] 利用众包的方式构建了一个小规模但是高质量的 Twitter 情感语料库。他们的情感

分类体系中包含 9 个类别：愤怒、期待、愉快、伤心、信任、惊奇、厌恶、恐惧及其它。在随机选取了 778 篇菲律宾语推文和 570 篇英语推文后，依托 CrowdFlower 平台对这些推文按照指定规范进行了人工标注，经过过滤后，获得 1 146 篇带情感标签的菲律宾语和英语推文。另外，他们的相关实验结果表明，现有的语言资源和工具还不足以对推文进行准确的情感分类。

通过以上分析可以看出，菲律宾语情感分析主要是有监督的、依赖人工标注的情感分类。而情感分类体系因不同学者而异，并且实验数据大多数是基于自己构建的小规模数据，因此无法客观地比较各种方法的效果。

2.6 命名实体识别

命名实体（name entity recognition，NER）是识别文本中具有特定意义的词语，如人名、地名等，并为其添加标注，它是自然语言处理的一个重要工具，对网络信息抽取、跨语言情感分析、机器翻译等上层应用起着非常重要的作用，对于语言研究工作也具有重要的支撑作用。但现有菲律宾语命名实体识别方面的研究成果还较少。

K. M. L. Eboña 等 [33] 利用最大熵法来实现菲律宾语小说摘录的命名实体识别。他们将命名实体分为人名、地名、机构名、日期、时间 5 类。其实验结果表明，基于 *F* 度量值，NERF−CRF（named entity recognizer Filipino text using conditional random field）的识别准确率达到 80.53%，其中在日期类别上的识别错误率为 0%，较差的是对地名和机构名的识别，错误率分别为 28.41% 和 13.10%。

与 K. M. L. Eboña 等 [33] 的研究成果相似，A. P. T. Alfonso 等 [34] 也提出了利用条件随机场实现菲律宾语文本命名实体识别系统 NERF−CRF。NERF−CRF 将命名实体分为人名、地名、日期、机构名 4 类。其实验结果表明，基于 *F* 度量值，NERF−CRF 的准确率达 83%，其中在日期类别上的识别错误率为 0%，较差的实体类别是地名和机构名，错误率分别为 42% 和 33%。

2.7 拼写检查

拼写检查旨在检索文本输入中因人为拼写错误导致的文本错误。现有拼写检查工具主要有 Microsoft Word 和 Google Docs，它们可以自动进行英语语法和拼写检查，并且提供修改建议，为语言学习者提供了极大的便利。诸如句法分析、树库、词性标注等工具，对于提高拼写检查效果有很大帮助 [35]。因此，菲律宾语拼写检查研究除了基于规则的方法外，有不少研究者开始考虑综合其他自然语言处理工具来提高纠错准确率。

E. D. Dimalen 等 [36] 实现了一个基于规则的菲律宾语拼写检查器，已经被作为插件整合在 OpenOffice 中，可用于检查拼写错误和语法错误。

N. Oco 等 [37] 利用 Language Tool，设计了一个基于词典及规则的拼写检查器，主要用于检查词语拼写错误、语法错误、漏词等情况。在 272 个带有错误的句子上进行测试，得知其准确率为 83%。

M. P. Go 等 [38] 也设计并实现了一个菲律宾语拼写检查器 Gramatika。他们先从高质量文本中获取 *n* 元模型、词性及词干信息，再利用这些信息学习出混合 *n* 元模型，最后通过学习出的模型和预定义的规则侦测文本中的拼写和语法错误，并给出修改建议。实验结果表明，该系统在错误表达上给出修改建议的准确率为 64%（248 个带有错误的句子），有 15% 的句子被判断为有错误（1 284 个没有错误的句子）。

由于菲律宾语语言资源及高效准确可用的语言分析工具的缺乏，与英语相比，菲律宾语的拼写检查研究较为滞后。N. L. Tsao 等 [39] 及 Huang C. C. 等 [40] 通过实验表明，POS 的引入使得拼写检查性能显著性提升。考虑到菲律宾语形态变化丰富，因此本课题组更加认为提高菲律宾语拼写检查效果，高质量的 POS 模型必不可少。

2.8 语料库构建

在人工构建菲律宾语语言资源 (例如词典、形态信息、语法规则库和语料库等) 方面的研究已经取得了很大进展。除此以外，由于人工构建语料库的内在困难，不少学者开始研究自动抽取高质量语言资源的技术。

E. P. Tiu 等 [41] 提出了一种从可比语料中自动提取双语词典的方法，其中英语为源语言，菲律宾语为目标语言。他们结合上下文抽取、聚类技术，并使用词性标签来定义单词的不同含义。实验结果表明，较前人研究的成果，他们获得的整体 *F* 值从 7.32% 提高到了 10.65%。

S. Dita 等 [42] 初步通过人工构建菲律宾国家语言的在线语料库，包括菲律宾语、宿雾语、伊洛卡诺语、希利盖农语和菲律宾手语。前 4 种语言包含 250 000 个单词的文本，而菲律宾手语包含 7 000 个视频。该在线语料库还提供了用于语言分析的自动化工具，例如字数统计。该项目后续考虑了自动获取文本、语音、视频等多模态语料资源。

文献 [42] 的工作是为德拉萨大学语言技术中心研发英菲机器翻译系统服务 [43]。除此以外，面对有

限的菲律宾语语言资源，基于菲律宾语语言委员会提供的词典，他们还构建了一个英菲词典，包含词条的形态学信息如词性标签等，具体如表 1 所示。

表 1　英菲词典 –DLSU

Table 1　English-Filipino dictionary (DLSU)

Database	Size
English-Filipino Entries	23 520
English-Filipino Attributes	7 762
Filipino-English Entries	20 540
Filipino-English Attributes	1 208

A. Borra 等 [44] 讨论了菲律宾语 Word Net 的构建，探讨了菲律宾语的形态用于构建分析器和生成器，以支持 Word Net 中的词干以及词缀序列对的收集。J. P. Ilao 等 [45] 针对搜索引擎如雅虎等，提出基于查询的方法来自动收集诸如新闻、博客评论等相关文本 (包含单语文本和双语文本)，并构建了语料库 Web Miner 系统。Web Miner 系统共收集了 14 600 个英菲平行句对，包含约 582 000 个菲律宾语单词。由于该系统不仅爬取新闻报道，还收集社交平台的评论等资源，因此获取的单语菲律宾语料库并不是完全正确的，包含拼写错误、语法错误、句子成分替换等问题。

A. El-Kishky 等 [46] 应用 URL（Uniform Resource Location）匹配规则，从 commoncrawl 语料库（http://commoncrawl.org/）中爬取高质量的跨语言文档数据集，包含 92 种不同语言 (含菲律宾语、印地语、德语等) 与英语对齐的文档对。他们首先使用人工注释来直接评估该数据集的质量，而后通过评估下游任务，即利用该对齐语料训练的机器翻译模型质量，进一步评估该数据集的质量。

R. A. Sagum 等 [47] 提出了基于决策树和 *n*-gram 模型的半监督方法来构建菲律宾语的语义知识库 FilWordNet。并将模型在 500 篇文档 (包含 25 618 个单词，其中含 15 377 个菲律宾语单语单词) 上测试，正确提取词干且进行 POS 的准确率高达 86.29%。

3　面临的问题

总体来说，在菲律宾语自然语言处理领域，语言资源不足，与英语、汉语等语种的自然语言处理研究相比，还存在较大差距。现有研究比较宽泛但不深入，在词法分析、句法分析、语义分析等底层技术及机器翻译、情感分析、拼写检查等应用技术中都有一些成果。其中，机器翻译的研究取得了较快的进展，拼写检查次之，而在句法分析、语义分析、命名实体识别等方面的研究成果相对较少。菲律宾语的机器翻译几乎都是涉及英语 – 菲律宾语的翻译，没有涵盖其他语言。这与菲律宾国家的语言政策有关，菲律宾国家的第二官方语言是英语，菲律宾政府和学术研究机构在英语和菲律宾语的语料构建及英菲机器翻译上投入了较多的人力和物力。而菲律宾语与其他语言对照的平行语料缺乏，研究投入不足。

虽然菲律宾语在自动构建语料库方面的研究取得了一定的进展，但是相较于英语、汉语等通用语种，菲律宾语仍然属于语言资源较为缺乏的低资源语言。大部分语料库构建研究旨在收集英菲平行句对或词对，主要服务于机器翻译；而关于自然语言处理其他领域的语料资源构建研究非常少。由于深度学习算法高度依赖于高质量、大规模的标注语料，导致无法有效运用深度学习方法于词法分析、句法分析、命名实体识别等方面。

在信息大爆炸时代，信息的精炼和提取成为一个重要的研究课题，而文本自动摘要是解决信息爆炸问题的关键技术之一，跨语言自动摘要技术可以让人们快速地了解不同国家和地区的信息。然而，根据已有文献调查发现，目前菲律宾语文本自动摘要方面的研究几乎为空白。

4　未来的发展方向

综合以上对菲律宾语自然语言处理现状分析可以得知，英语 – 菲律宾语平行语料较为丰富，有力地推动了机器翻译的研究进展。面对丰富的英语 – 菲律宾语平行语料，如何通过跨语言处理技术，构建汉语 – 菲律宾语平行语料库，成为我国研究汉语 – 菲律宾语机器翻译、跨语言自动摘要等任务的首要解决问题。

针对菲律宾语的其他自然语言处理领域语料匮乏的问题，同时在词法分析、句法分析、语义分析等任务上无法使用海量无标注语料进行深度学习等，十分必要构建相关领域较大规模、开放的标注数据库。面对资源缺乏的基础问题，尽管菲律宾语形态变化丰富，但只要总结足够多的形态规则就可以构建形态学信息语料库；而正确的形态学信息可为词性标注和句法分析等提供重要的语言特征，有利于提高其他自然语言处理任务的性能，从而利用半监督的资源构建技术促进其他领域语言资源的构建。

在大规模、高质量、开放的语言资源构建的前提下，深度学习应用于菲律宾语自然语言处理的方法研究成为可能。在基本理论和模型创新的基础上，鉴于菲律宾语的句子语法结构较为灵活，并结合基于规

则、基于统计和深度学习的方法，可在一定程度上解决由菲律宾语复杂的语言特征造成的诸如词义多样、句法结构歧义等问题，从而推动命名实体识别、句法分析、语法纠错、知识图谱构建以及语义分析等方面的研究。

最后，考虑到信息爆炸时代下文本自动摘要技术的重要性，可借鉴其他语言的文本自动摘要研究技术，探讨基于规则、基于图模型、基于结构等方法对菲律宾语文本自动摘要的适用性，以填补菲律宾语自动文摘研究的空缺，这也是未来研究的重要方向。

参考文献：

[1] FORTES F. A Constraint-Based Morphological Analyzer for Concatenative and Nonconcatenative Morphology of Tagalog Verbs[D]. Manila：De La Salle University Manila，2002.

[2] FORTES-GALVAN F C，ROXAS R E. A Constraint-Based Morphological Analyzer for Concatenative and Non-Concatenative Morphology[C]//Proceedings of the 20th Pacific Asia Conference on Language，Information and Computation. Wuhan：Chinese Information Processing Society of China，2006：273-279.

[3] ROXAS R，MULA G. A Morphological Analyzer for Filipino Verbs[C]//Proceedings of the 22nd Pacific Asia Conference on Language，Information and Computation. Manila：De La Salle University，2008：467-473.

[4] BONUS D E. The Tagalog Stemming Algorithms (TagSA)[C]//Proceedings of the Natural Language Processing Research Symposium. Manila：De La Salle University，2003：63-67.

[5] BAUMANN P，PIERREHUMBERT J B. Using Resource-Rich Languages to Improve Morphological Analysis of Under-Resourced Languages[C]//Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC). Reykjavik：European Language Resources Association，2014：3355-3359.

[6] CHENG C K，RABO V S. TPOST：A Template-Based，*n*-Gram Part-of-Speech Tagger for Tagalog[J/OL]. Journal Research in Science，Computing and Engineering (JRSCE)，2004，3(1). [2019-12-30]. http://xsite.dlsu.edu.ph/research/ centers/adric/nlp/jrsce-tpost.pdf.

[7] ERLYN M，YUJI M. Factors Affecting Part-of-Speech Tagging for Tagalog[C]//Proceedings of the 23rd Pacific Asia Conference on Language，Information and Computation. Hong Kong：City University of Hong Kong，2009：763-770.

[8] REYES C D E，SUBA K R S，RAZON A R，et al. SVPOST：A Part-of-Speech Tagger for Tagalog Using Support Vector Machines[C/OL]//Proceedings of the 11th Philippine Computing Science Congress. Naga City：Ateneo de Naga University，2011. [2019-12-30]. https://www.researchgate. net/publication/260389578.

[9] NOCON N，BORRA A. SMTPOST Using Statistical Machine Translation Approach in Filipino Part-of-Speech Tagging[C]//Proceedings of the 30th Pacific Asia Conference on Language，Information and Computation: Posters. Seoul：[s. n.]，2016：391-396.

[10] GO M P，NOCON N. Using Stanford Part-of-Speech Tagger for the Morphologically-Rich Filipino Language[C]//Proceedings of the 31st Pacific Asia Conference on Language，Information and Computation. Manila：The National University，2017：81-88.

[11] OLIVO J F T，HARI P J T，DELA FUENTE M B. CRFPOST：Part-of-Speech Tagger for Filipino Texts Using Conditional Random Fields[C]//Proceedings of the 2019 2nd International Conference on Algorithms，Computing and Artificial Intelligence，Sanya China. New York：ACM，2019：444-449.

[12] CLARK A. Unsupervised Induction of Stochastic Context-Free Grammars Using Distributional Clustering[C]//Proceedings of the 2001 Workshop on Computational Natural Language Learning-Volume 7. Morristown：Association for Computational Linguistics，2001：13.

[13] ALCANTARA D L，BORRA A. Constituent Structure for Filipino：Induction Through Probabilistic Approaches[C]// Proceedings of the 22nd Pacific Asia Conference on Language，Information and Computation. Manila：De La Salle University，2008：113-122.

[14] MANGUILIMOTAN E，MATSUMOTO Y. Dependency-Based Analysis for Tagalog Sentences[C]//Proceedings of the 25th Pacific Asia Conference on Language，Information and Computation. Singapore：Institute of Digital Enhancement of Cognitive Processing，Waseda University，2011：343-352.

[15] DOMINGO E，ROXAS R E. Utilizing Clues in Syntactic Relationship for Automatic Target Word Sense Disambiguation[J]. Journal of Research in Science，Computing and Engineering (JRSCE)，2006，3(3)：18-24.

[16] MISTICA M，BALDWIN T. Recognising the Predicate-Argument Structure of Tagalog[C]//Proceedings of Human Language Technologies：The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics，Companion Volume：Short Papers. Morristown：Association for Computational Linguistics，2009：257-260.

[17] BERGSMA S，BHARGAVA A，HE H，et al. Predicting the Semantic Compositionality of Prefix Verbs[C]//

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Massachusetts：Association for Computational Linguistics，2010：293−303.

[18] ANDREI A L. Development and Evaluation of Tagalog Linguistic Inquiry and Word Count (LIWC) Dictionaries for Negative and Positive Emotion[EB/OL]. [2019−12−30]. https://www.mitre.org/sites/default/files/publications/pr_14-3858-development-evaluation-of-tagalog-linguistic-inquiry.pdf.

[19] ROXAS R，SANCHEZ W，BUENAVENTURA M. Final Report of Machine Translation from English to Filipino：Second Phase[R]. Manila：Department of Science and Technology，Philippine Council for Advanced Science and Technology Research and Development，1999：1−3.

[20] BORRA A. A Transfer-Based Engine for an English to Filipino Machine Translation Software[D]. Los Baños：University of the Philippines Los Baños，1999.

[21] BORRA A，CHAN E A O，LIM C I R，et al. LFG-Based Machine Translation Engine for English and Filipino[C]//4th National Natural Language Processing Research Symposium. Manila：De La Salle University，2007：36−42.

[22] ALLMAN T，BEALE S，DENTON R. Toward an Optimal Multilingual Natural Language Generator：Deep Source Analysis and Shallow Target Analysis[J]. Philippine Computing Journal，2014，9(1)：55−63.

[23] ROXAS R. A Hybrid English-Filipino Machine Translation System[C]//3rd National Natural Language Processing Research Symposium. Manila：De La Salle University，2006：1−4.

[24] ROXAS R E O，BORRA A，CHENG C K，et al. Building Language Resources for a Multi-Engine English-Filipino Machine Translation System[J]. Language Resources and Evaluation，2008，42(2)：183−195.

[25] ONG E，GO K，NUÑEZ V A，et al. Template-Based English-Filipino Machine Translation System[C]// Proceedings of the 4th National Natural Language Processing Research Symposium. Manila：De La Salle University，2007：43−47.

[26] ANG J，CHAN M R，GENATO J P，et al. Development of a Filipino-to-English Bidirectional Statistical Machine Translation System that Dynamically Updates Via User Feedback[C/OL]//Proceedings of the 12th International Workshop on Spoken Language Translation, Da Nang, Vietnam. [2019−12−30]. http://workshop2015.iwslt.org/downloads/IWSLT_2015_RP_6.pdf.

[27] TACORDA A J，IGNACIO M J，OCO N，et al. Controlling Byte Pair Encoding for Neural Machine Translation[C]//2017 International Conference on Asian Language Processing (IALP). Singapore：IEEE，2017：168−171.

[28] LAZARO A N，OCO N，ROXAS R E. Developing a Bidirectional Ilocano-English Translator for the Travel Domain：Using Domain Adaptation Techniques on Religious Parallel Corpora[C]//11th International Conference of the Asian Association for Lexicography. Guangzhou：Guangdong University of Foreign Studies，2017：889.

[29] REGALADO R V J，CHUA J L，CO J L，et al. Subjectivity Classification of Filipino Text with Features Based on Term Frequency：Inverse Document Frequency[C]//2013 International Conference on Asian Language Processing. Urumqi：IEEE，2013：113−116.

[30] PIPPIN M，ODASCO R，DE JESUS R，et al. Classifications of Emotion Expressed by Filipinos Through Tweets[C]// Proceedings of the International Multi Conference of Engineers and Computer Scientists. Hong Kong：[s. n.]，2015：18−20.

[31] PATACSIL F，FERNANDEZ P. Blog Comments Sentiment Analysis for Estimating Filipino ISP Customer Satisfaction[C]//Proceedings of the IRES 11th International Conference. Bangkok：[s. n.]，2015：77−84.

[32] LAPITAN F R，BATISTA-NAVARRO R T，ALBACEA E. Crowdsourcing-Based Annotation of Emotions in Filipino and English Tweets[C]//Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016). Osaka：The COLING 2016 Organizing Committee，2016：74−82.

[33] EBOÑA K M L，LLORCA JR O S，PEREZ G P，et al. Named-Entity Recognizer (NER) for Filipino Novel Excerpts Using Maximum Entropy Approach[J]. Journal of Industrial and Intelligent Information，2013，1(1)：63−67.

[34] ALFONSO A P T，DOMINGO I V R，GALOPE M J F，et al. Named Entity Recognizer for Filipino Text Using Conditional Random Field[J]. International Journal of Future Computer and Communication，2013，2(5)：376.

[35] ROZOVSKAYA A，ROTH D. Building a State-of-the-Art Grammatical Error Correction System[J]. Transactions of the Association for Computational Linguistics，2014，2：419−434.

[36] DIMALEN E D，DIMALEN D M D. An Open Office Spelling and Grammar Checker Add-in Using an Open Source External Engine as Resource Manager and Parser[C]// Proceedings of the 4th National Natural Language Processing Research Symposium (NNLPRS). Manila：De La Salle University，CSB Hotel，2007：69−73.

[37] OCO N，BORRA A. A Grammar Checker for Tagalog

Using Language Tool[C]//Proceedings of the 9th Workshop on Asian Language Resources. Chiang Mai：Asian Federation of Natural Language Processing，2011：2-9.

[38] GO M P，NOCON N，BORRA A. Gramatika：A Grammar Checker for the Low-Resourced Filipino Language[C]//TENCON 2017-2017 IEEE Region 10 Conference. Penang：IEEE，2017：471-475.

[39] TSAO N L，WIBLE D. A Method for Unsupervised Broad-Coverage Lexical Error Detection and Correction[C]//Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications. Morristown：Association for Computational Linguistics，2009：51-54.

[40] HUANG C C，CHEN M H，HUANG S T，et al. EdIt：A Broad-Coverage Grammar Checker Using Pattern Grammar[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics：Human Language Technologies：Systems Demonstrations. Morristown：Association for Computational Linguistics，2011：26-31.

[41] TIU E P，ROXAS R E. Automatic Bilingual Lexicon Extraction for a Minority Target Language[C]//Proceedings of the 22nd Pacific Asia Conference on Language，Information and Computation. Manila：De La Salle University，2008：368-376.

[42] DITA S，ROXAS R E，INVENTADO P. Building Online Corpora of Philippine Languages[C]//Proceedings of the 23rd Pacific Asia Conference on Language，Information and Computation. Hong Kong：City University of Hong Kong，2009：646-653.

[43] CHU S. Language Resource Development at DLSU-NLP Lab[C/OL]//The School of Asian Applied Natural Language Processing for Linguistics Diversity and Language Resource Development ADD-4：Language Resource Technology. Bangkok：[s. n.]，2009. [2019-12-30]. http://xsite.dlsu.edu. ph/research/centers/adric/nlp/downloads/(ADD4)%20Shirley-Language%20Resource%20Development%20at%20DLSU-NLP%20Lab.pdf.

[44] BORRA A，PEASE A，ROXAS R，et al. Introducing Filipino WordNet[C/OL]//Principles，Construction and Application of Multilingual Wordnets：Proceedings of the 5th Global WordNet Conference. Mumbai：[s. n.]，2010. [2019-12-30]. https://pdfs.semanticscholar.org/3b5a/e33f2e5b774bc985d639c308abd3cbdf759f.pdf?_ga=2.212162516.1992774683.1586851367-107277088.1583911469.

[45] ILAO J P，GUEVARA R C L. Mining Filipino-English Corpora from the Web[C]//International Symposium on Multimedia and Communication Technology (ISMAC2010). Manila：[s. n.]，2010：8-10.

[46] EL-KISHKY A，CHAUDHARY V，GUZMAN F，et al. A Massive Collection of Cross-Lingual Web-Document Pairs[EB/OL]. [2019-12-30]. https://arxiv.org/abs/1911.06154.

[47] SAGUM R A，RAMOS A D，LLANES M T. FICOBU：Filipino WordNet Construction Using Decision Tree and Language Modeling[J]. International Journal of Machine Learning and Computing，2019，9(1)：103-107.

（责任编辑：廖友媛）