doi:10.3969/j.issn.1673−9833.2020.03.001

融合 Self-Attention 机制和 *n*-gram 卷积核的印尼语复合名词自动识别方法研究

丘心颖[1,2]，陈汉武[1,2]，陈　源[1,2]，谭立聪[1,2]，张　皓[1,2]，肖莉娴[3]

（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；
3. 广东外语外贸大学 东方语言文化学院，广东 广州 510006）

摘　要：针对印尼语复合名词短语自动识别，提出一种融合 Self-Attention 机制、*n*-gram 卷积核的神经网络和统计模型相结合的方法，改进现有的多词表达抽取模型。在现有 SHOMA 模型的基础上，使用多层 CNN 和 Self-Attention 机制进行改进。对 Universal Dependencies 公开的印尼语数据进行复合名词短语自动识别的对比实验，结果表明：TextCNN+Self-Attention+CRF 模型取得 32.20 的短语多词识别 F_1 值和 32.34 的短语单字识别 F_1 值，比 SHOMA 模型分别提升了 4.93% 和 3.04%。

关键词：印尼语复合名词短语；Self-Attention 机制；卷积神经网络；自动识别；条件随机场

中图分类号：TP31　　**文献标志码**：A　　**文章编号**：1673−9833(2020)03−0001−09

引文格式：丘心颖，陈汉武，陈　源，等 . 融合 Self-Attention 机制和 *n*-gram 卷积核的印尼语复合名词自动识别方法研究 [J]. 湖南工业大学学报，2020，34(3)：1−9.

Automatic Recognition of Indonesian Compound Noun Phrases with a Combination of Self-Attention Mechanism and *n*-gram Convolution Kernel

QIU Xinying[1,2]，CHEN Hanwu[1,2]，CHEN Yuan[1,2]，TAN Licong[1,2]，ZHANG Hao[1,2]，XIAO Lixian[3]

（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；3. Faculty of Asian Languages and Cultures，Guangdong University of Foreign Studies，Guangzhou 510006，China）

Abstract：In view of the automatic recognition of Indonesian compound noun phrases, this paper proposes a method with Self-Attention mechanism, *n*-gram convolution kernel neural network and statistical model combined together so as to improve the performance of the existing multi-word expression extraction model. On the basis of the existing SHOMA model, a further improvement can be made by using the multi-layer CNN and Self-Attention mechanism, followed by an automatic recognition of compound noun phrases based on Indonesian data disclosed by Universal Dependencies. The comparative experiment results show that the F_1 multi-word phrase recognition value of 32.20, as well as the F_1 single-word recognition value of 32.34 obtained by TextCNN+Self-Attention+CRF model obtains respectively is 4.93% and 3.04% respectively higher than that of SHOMA model.

收稿日期：2020−03−29

基金项目：广东省教育厅特色创新基金资助项目（2015KTSCX033），国家社会科学基金资助项目（17BGL068）

作者简介：丘心颖（1970−），女，广东梅州人，广东外语外贸大学教授，主要从事计算语言学和信息检索方面的教学与研究，E-mail：xy.qiu@foxmail.com

Keywords：Indonesian compound noun phrase；Self-Attention mechanism；convolutional neural network；automatic identification；conditional random field

1 研究背景

MWEs（multiword expressions）是一种由两个或者两个以上词汇所组成的语义单元。它们作为多种语言中普遍存在的特殊语言形式，其语义属性以及释义不能简单地由其构成的词汇得出。由于在句法以及语义特征上的特性，多词表达是自然语言处理的一大难题，尤其对于句法分析和机器翻译尤为关键。

复合名词是多词表达的一类重要形式，是普遍存在于各种语言中的一种特殊而又常见的语言结构。简单来说，复合名词短语就是由两个或者两个以上名词构成的名词短语。据祝慧佳 [1] 对国内外相关工作的总结以及对 HIT-IR（Harbin Institute of Technology-Information Retrieve）汉语依存关系树库的统计，在英文中，包括小说体散文、新闻以及科技摘要等多种文体在内，存在大量复合名词，并且其数量和种类都在增长 [2-4]；HIT-IR 关系树库内的 10 000 句语料中也发现了大量的复合名词。陈昌熊 [5] 指出，复合词在大多数的技术说明书中非常普遍，通常一个科技术语本身就是一个复合词。高年华 [6] 指出，印尼语的复合名词依据其组成方式可主要分为 3 类具体如表 1 所示。印尼语复合名词自动化抽取的研究可以应用于多种场景，包括印尼语的新词发现、词典自动扩充、机器翻译、印尼语教学以及句法分析等。

表 1 复合名词类别举例

Table 1 Examples of Indonesian compound nouns

复合名词类别	复合词	拆分词
由名词和名词组成	labu air 冬瓜	labu 瓜，air 水
	rumah sakit 医院	rumah 房子，sakit 病
	orang hutan 猩猩	orang 人，hutan 森林
由名词和动词组成	kertas ampelas 砂纸	kertas 纸，ampelas 擦
	kapal terbang 飞机	kapal 船，terbang 飞
	lampu témpél 壁灯	lampu 灯，témpél 粘贴
由名词和形容词组成	labu merah 南瓜	labu 瓜，merah 红
	emas padu 纯金	emas 金，padu 纯的
	manis mulut 甜言蜜语	manis 甜，mulut 口

多词表达研究，尤其是复合名词的识别，具有重要的理论价值和应用前景，并且在英语、汉语等通用语种中的相关研究已经取得了一定的成效 [7-11]。目前主要是以大量的语料数据库、词典、复合名词语义知识库、依存关系树库等作为驱动，通过基于规则、基于统计、基于神经网络的方法或以上方法的结合，以实现复合名词的自动识别。基于规则的多词表达技术，一般具体研究某一种多词表达类型或者某一特定领域，结合了语言学的知识，构造了描述语言的规则集合。基于统计的多词表达技术，是指从词频等可用于统计的信息出发，通过使用各种数学公式或其他度量方法来度量多词表达内部的结合程度，以及多词表达与上下文的结合程度等。随着神经网络的飞速发展，当前最新的技术大多是通过卷积神经网络（convolutional neural networks，CNN）、双向长短期记忆网络（bidirectional long short-term memory，Bi-LSTM）等神经网络结合条件随机场（condition random field，CRF）的方法。近年来，Self-Attention 机制通过捕获同一个句子中单词之间的一些句法特征或语义特征，更容易得到句子中长距离相互依赖的特征，对神经网络处理文本信息的效果有显著提升，从而在各种自然语言处理（natural language processing，NLP）任务中得到认可。

印尼语是一类典型的黏着语，具有丰富的形态变化。作为一种非通用语言，印尼语与汉语、英语等通用语种的语法规则和语义关系不尽相同，并且缺少相关的语料和词典，这导致其复合名词的识别面临着更大的挑战。

由于现有的研究技术和方法在印尼语复合名词识别任务中也并非是最有效的，因此本研究提出基于有限的标注语料，检验多词表达识别的最新神经网络算法对于印尼语复合名词自动识别的效果。课题组前期研究中注意到印尼语的复合名词大多数为二元词和三元词，少数为四元词，因此提出检验 *n*-gram 卷积核对于现有模型的影响。同时，考虑到 Self-Attention 处理序列信息的优势，提出以融合 Self-Attention 和 *n*-gram 卷积核进一步改进现有的 state-of-the-art 模型，即 SHOMA 模型 [11] 在印尼语复合名词识别中的应用。

2 相关研究

目前国内外对印尼语复合名词提取的相关研究尚未开展，但从大规模语料库中自动提取多词表达式、短语或搭配等语言知识的研究已广泛开展，并获得了较多的研究成果与进展。这些研究可以为印尼语复合名词提取的研究提供借鉴。

2.1 语言学研究

在印尼语的语言结构方面，高华年 [6] 从语言学的角度对印尼语的名词结构进行了深入研究，总结出印尼语复合名词主要有如下 3 类：由名词和名词组合而成、由名词和动词组合而成、由名词和形容词组合而成。每个类别下又根据复合名词中组成词的主从关系、词义关系等细分为多个不同的小类。其阐述了印尼语名词的构造规律，为印尼语复合名词的提取提供了语言学基础。

在复合名词短语的分类方面，刘鹏远等 [7] 总结了国外复合名词短语语义关系分类的研究，主要有两种路线：一种是通过复合短语内部各个成分的语义类来定义其语义关系 [12]，另一种则是基于删除谓词的语义类来定义复合名词短语内部成分的语义关系 [13-14]。B. Warren[13] 在对英语复合名词短语的研究中，通过删除谓词对而获得“*N*1+*N*2”复合名词短语，然后对其名词成分之间的语义关系进行分类，并根据可删除谓词的语义类别，提出了做修饰成分的名词和核心名词之间存在 12 种语义关系，为英语复合名词的提取和识别提供了词汇学和语义学基础。

2.2 多词表达提取研究

2.2.1 通用语种多词表达提取

从 20 世纪 90 年代开始就有学者对 MWEs 提取进行研究，如 F. Smadja[15] 设计了 Xtract 系统来提取词语搭配（collocations）。在基于语言规则与统计的多词表达提取方法方面，S. Piao 等 [16] 使用对数似然以及卡方的方法从中国电子信息产业发展研究院的中文语料库中抽取了中文 MWEs。在国际计算语言学会 2009 年主办的 MWEs 专题讨论会上，H. Wakaki 等 [17] 介绍了如何使用对数线性模型抽取日语的 MWEs。唐亮等 [18] 根据重复频次、左右邻接熵、内部关联度、多词嵌套等方法，在汉日平行语料库中抽取出多词短语。

另外，一些研究者通过引入语义信息来提高 MWEs 的识别效率。如 T. Baldwin[19] 和 G. Katz 等 [20] 使用向量空间计算语义距离的方法识别 MWEs，T. van de Cruys 等 [21] 在 2007 年使用聚类和优选语义的方法识别了 MWEs。肖健等 [22] 提出了一种基于语义模板与基于统计工具相结合的方法，该方法采用基于词表和分布的方法计算词语间的相似度，扩大了 MWEs 的覆盖范围，并且从三元组可比语料库中自动提取了本族英语 MWEs。梁颖红等 [23] 提出了半监督策略抽取汉语多词表达，并且在聚类算法的中后期加入有监督的信息，使分类器能使用正确的标注信息进行训练。

近年来，许多学者结合机器学习和深度学习方法实现了多词表达提取。如 J. R. Williams[8] 使用了非单词标记，即“边界”，提出了一种跨越 19 种不同的语言用于 MWEs 分割的监督式机器学习细粒度文本分块算法。M. J. Hosseini 等 [9] 提出了使用“2-CRF”（double-chained conditional random field）来进行英语语料多词表达的识别。O. Rohanian 等 [10] 结合 GCN（graph convolutional network）和 multi-head self-attention 两种神经网络结构用于多词表达的提取。其中，GCN 利用依赖性来解析文体信息，然而自注意力机制（self-attention）则关注长期关系。S. Taslimipoor 等 [11] 提出了一种适用于多词表达识别任务的深度学习体系结构，它是一个由卷积层和递归层组成的神经结构，并且在顶层增加了一个可选的条件随机场层，卷积层通过捕获输入序列的 *n*-gram 词汇信息，使得这个系统在 Parseme 共享任务中的表现明显优于其他所有参与的系统。

2.2.2 非通用语种多词表达提取

受通用语种的多词表达提取研究成果的启发，国内外许多学者在此基础上结合非通用语种的特点，提出了一些非通用语种多词表达提取的方法。已有研究结果表明，借鉴通用语种的多词表达提取方法，对非通用语种的多词表达提取大有帮助。其中赵维纳等 [24] 结合藏语三音动词短语的结构，利用统计算法和语言规则库进行过滤，提出了一种统计和规则相结合的藏语三音动词短语的自动抽取算法。麦热哈巴·艾力等 [25] 讨论了目前常见的互信息、对数似然比和卡方 3 种统计方法在维吾尔语多词表达抽取方面的影响。张海军 [26] 总结了近年来维吾尔语短语识别的有关语言学研究成果，重点梳理了维吾尔语短语自动抽取的相关研究方法。古丽扎达·海沙等 [27] 提出了一种搭配规则集与最大熵相结合的混合策略方法对哈萨克语 KzBaseVP（基本动词短语）进行识别，取得了较好的实验结果。

3 研究方法和评价指标

3.1 研究内容

本研究在现有多词表达识别模型的基础上进行改进，采用神经网络和统计方法相结合的模型，探究其对印尼语复合名词的自动发现。用神经网络表示学习印尼语文本的特征，用概率模型表示学习到的特征，以固定的输出方式得到对复合名词短语的预测结果。通过建立单个模型和组合模型来对以上两个研究问题分别进行探索，并分析实验结果，得出相关结论。本

研究以基于深度学习方法的印尼语复合名词短语的自动识别为目的，提出如下 2 个研究问题（research question，RQ）。

RQ1 *n*-gram 卷积核的机制对多词表达的识别是否有效？

研究问题 1（RQ1）的提出是基于对印尼语复合名词 *n*-gram 数据的观察，图 1 展示的是复合名词短语的 *n*-gram 分布情况。

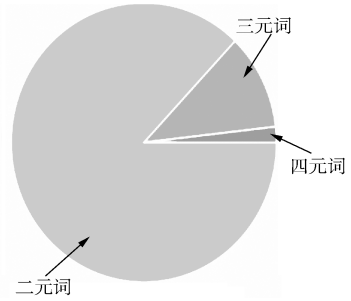

图 1 *n* 元复合名词短语分布情况

Fig. 1 Distribution of *n*-gram compound noun phrases

由于复合名词包含二元、三元、四元词汇，因而提出比较 *n*-gram 卷积核和基准模型以及无 *n*-gram 卷积核模型的识别效果。

RQ2 Self-Attention 机制是否可以对现有的神经网络方法带来改善？

现有的多词表达 state-of-the-art 模型（即 SHOMA[11]）采用了 Bi-LSTM（Bidirectional LSTM）模型 [28]。由于 Self-Attention 每个节点都可以捕获到序列上其他节点的信息，可以用于学习序列中复合名词短语的表示结构，故本研究提出融合 Self-Attention 对多词表达识别的影响。

3.2 研究框架

采用 BIO 标签将印尼语语料进行复合名词短语标注，之后转化为词嵌入表示作为不同模型的输入。根据模型结构的差异，用于探讨本文提出的 2 个研究问题，即分别评价 *n*-gram 卷积特征对模型的影响以及 Self-Attention 机制和 *n*-gram 卷积特征对模型的影响。研究框架的具体流程如图 2 所示。

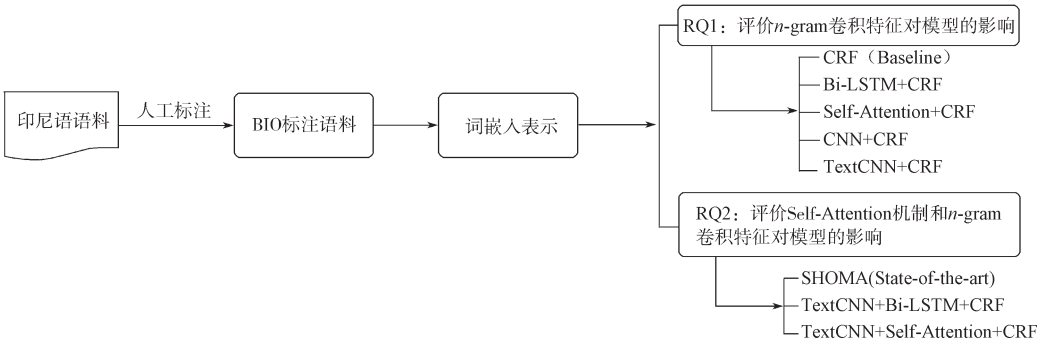

图 2 研究框架流程

Fig. 2 Research framework

3.2.1 基准模型

多词表达发现的方法有基于规则的方法、基于统计的方法和基于神经网络的方法。本文采用基于统计的 CRF 模型作为基准模型。

CRF 是自然语言处理领域的常用算法之一，常被用于句法分析、命名实体识别、词性标注等。在给定输入序列和输出序列的情况下，CRF 可以通过学习输出序列的表示形式，来约束输入序列经过计算之后得到的输出序列的格式。

本研究中使用单个 CRF 层作为基准模型，将印尼语文本学习到的词嵌入向量和文本对应的标注序列作为 CRF 层的输入，得到的输出即是预测的标注序列。

3.2.2 State-of-the-art 模型

文献 [11] 提出的 SHOMA 模型（见图 3 中的 SHOMA），是一种基于 CNN 和 Bi-LSTM 的神经网络同时结合 CRF 统计方法的模型，可用于 20 种语言多词表达的抽取。

SHOMA 模型在罗马尼亚语上取得最佳 F_1 值，为 87.18，在部分非通用语种上也表现出较好的效果。表明该模型可以被复用于通用语种和非通用语种的多词表达识别。

3.3 融合 Self-Attention 和 *n*-gram 卷积核的模型

鉴于印尼语的复合名词大多数为二元词和三元词，少数为四元词，本文首先采用 2 种形式的 CNN 用于捕获 *n*-gram 信息，检验 *n*-gram 特征对于神经网络模型的影响：

1）使用卷积核大小为 3 的单层 CNN（见图 4），仅提取句子中 3-gram 的词汇信息。实验中也尝试了卷积核大小为 2 和 4 的单层 CNN，但效果没有卷积核大小为 3 的好。

2）使用卷积核大小为 2，3，4 的三层 CNN（见

图 5），即采用 TextCNN[29] 的思想，可用于同时提取 2-gram、3-gram 和 4-gram 的词汇信息。

在本实验中，同时使用 TextCNN 对 SHOMA 的模型进行改进，即在原本的 CNN 层中，添加一层卷积核大小为 4 的 CNN 层，使得模型可以捕获到 4-gram 的词汇信息（如图 3 中的 TextCNN+Bi-LSTM+CRF 所示）。

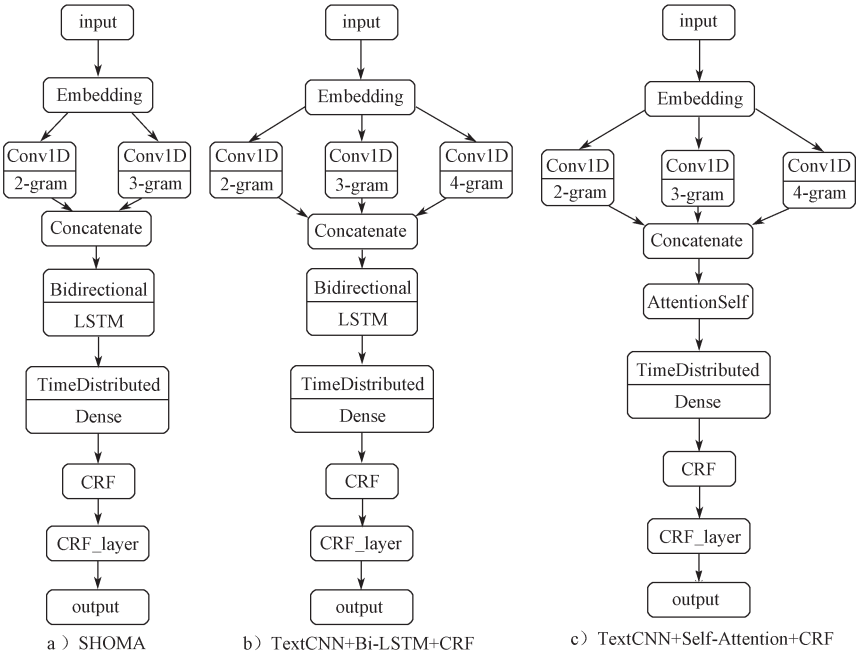

图 3　模型架构比较

Fig. 3　Model architecture comparison

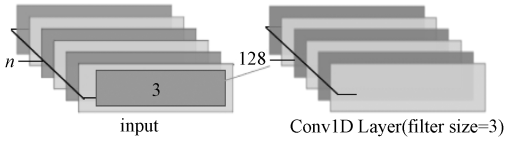

图 4　单层 CNN 模型

Fig. 4　Single layer CNN model

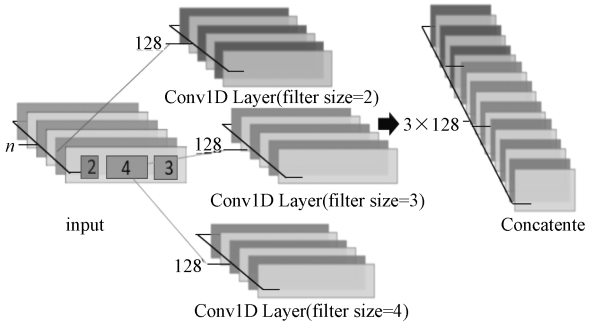

图 5　三层 CNN 模型

Fig. 5　Three-layer CNN model

Self-Attention 将输入向量 $\boldsymbol{X}$（分别乘以 $\boldsymbol{W}_{\mathrm{Q}}$，$\boldsymbol{W}_{\mathrm{K}}$，$\boldsymbol{W}_{\mathrm{V}}$ 权值矩阵得到 3 个向量 Query($\boldsymbol{Q}$)，Key($\boldsymbol{K}$)，Value($\boldsymbol{V}$)，并对 $\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$ 进行如图 6 所示的 Scaled Dot-Product Attention，让每个输入节点的 $\boldsymbol{V}$ 都能捕获到其他节点的 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$。由于这种机制使得 Self-Attention 每个输出节点都会保留序列上所有输入节点的信息，可以捕获长距离依赖的关系。在处理序列信息时，Bi-LSTM 能捕获到双向语义依赖，但是其效率明显低于 Self-Attention 机制的。本文使用 Self-Attention 替换 Bi-LSTM 的方法（如图 3 中的 TextCNN+Self-Attention+CRF 模型所示），探讨融合 Self-Attention 机制对印尼语复合名词短语自动识别效果的影响。

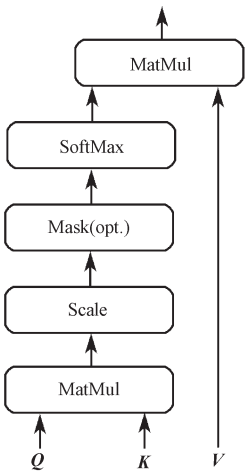

图 6　Scaled Dot-Product Attention 结构

Fig. 6　Scaled Dot-Product Attention structure

CNN 和 Self-Attention 能够有效地捕获上下文特征，但它们对最终输出标签的结构一无所知。为了有效地预测序列的标签，模型除了要学习数据的特征

外，还要学习输出的结构。

综上所述，实验模型结合 3 层 CNN 模型，并采用 CRF 作为最终的输出预测层，学习已知的标注序列，将 Self-Attention 学习到的特征按照约束输出预测的标注序列，实现印尼语复合名词多词表达的自动识别。

3.4 评价指标

实验中使用精度（precision）、召回率（recall）和 F_1 值（F_1-score）作为评价指标 [30]，在两种情况下对结果进行评估：一种情况是严格匹配（基于多词识别），这种情况下所有的多词表达的组件都被视为一个单元，应该被正确区分；另一种情况是模糊匹配（基于单字识别），它计算的是预测的单词及其对应的标注。

3.4.1 基于短语多词识别的 F_1 值

C_N 表示正确抽取出来的复合名词短语的数量，P_N 表示总共抽取出来的复合名词短语的数量，T_N 表示所有测试数据中复合名词短语的数量。基于短语多词识别的评价指标如表 2 所示。

表 2 基于短语多词识别的评价指标

Table 2 Evaluation metric based on multi-word recognition

评价指标	计算公式
准确率 P	$P = C_N / P_N$
召回率 R	$R = C_N / T_N$
F_1 值	$F_1 = 2PR / (P+R)$

3.4.2 基于短语单字识别的 F_1 值

基于单字识别的评价指标如表 3 所示。

表 3 基于单字识别的评价指标

Table 3 Evaluation index based on single-word recognition

评价指标	计算公式
准确率 P	$P = C_W / P_W$
召回率 R	$R = C_W / T_W$
F_1 值	$F_1 = 2PR / (P+R)$

其中，C_W 表示正确识别的复合名词短语中单词的数量，P_W 表示总共抽取出来的复合名词短语的单词的数量，T_W 表示所有测试数据中复合名词短语的单词的数量。

4 实验设置

4.1 数据准备

本文的数据来自 Universal Dependencies 公开的印尼语标准数据集，包含了 1 000 句印尼语文本数据，其中每一句印尼语都有对应的词性标注。

为了获取标签序列，根据词性标注对该数据集使用 BIO 标注格式（见表 4），标记出复合名词短语，得到了 332 个复合名词短语，其中二元词 289 个，三元词 37 个，四元词 6 个。

表 4 BIO 标注

Table 4 BIO annotation

标签名称	标 签 意 义
B（begin）	表示复合名词短语的开始单词
I（inside）	表示复合名词短语的其他单词
O（other）	表示与复合名词短语无关的单词

为了去除数据中一些无关的信息，增强复合名词短语在语句中的表示，对原始数据进行如下处理：

1）将原始数据和标签序列一一对应存放，并按句划分；

2）替代特殊符号，减少噪声的出现；

3）根据人工标注的结果，使用 BIO 格式对完成预处理的数据进行标注，让模型学习到输出序列的表示。

实验结果表明，使用 BIO 标注格式，并将预测结果是单个词语的删除，对印尼语复合名词短语的自动识别是有效的。

印尼语语料通过处理后的样例如表 5 所示。从表 5 中的标注结果可以得知，transisi media sosial 是一个复合名词短语，将其提取词条后转化为复合名词识别标签序列。为了使得印尼语文本转化为模型输入的格式，本文采用随机训练的方法，用 one-hot 表示印尼语文本，将 one-hot 向量输入到神经网络模型中进行端到端的训练。

表 5 数据预处理和标注举例

Table 5 Illustration of data preprocessing and annotation

原句	Bagi mereka yang mengikuti transisi media sosial di Capitol Hill.									
标注结果	Bagi	mereka	yang	mengikuti	transisi	media	sosial	di	Capitol	Hill
	O	O	O	O	B	I	I	O	O	O

4.2 参数设置

为了研究基于深度学习方法的印尼语复合名词短语的自动识别，设置多个模型来验证前文所提出的 2 个研究问题。这些模型训练时所设置的部分参数如表 6 所示。

在构建模型时，使用 relu 作为模型的激活函数。为了提取完整的 *n*-gram 词汇信息，CNN 层没有采用 dropout；在 Bi-LSTM 层中，使用 0.5 的 dropout 和 0.2 的 recurrent dropout。

表 6　模型参数设置

Table 6　Model parameters

模　型	Bi-LSTM 神经元数量	Self-Attention 神经元数量	卷积核大小	卷积核数量
Bi-LSTM+CRF	300			
Self-Attention+CRF		256		
CNN+CRF			3	128
TextCNN+CRF			2, 3, 4	128
SHOMA	300		2, 3	128
TextCNN+Bi-LSTM+CRF	300		2, 3, 4	128
TextCNN+Self-Attention+CRF		256	2, 3, 4	128

在模型训练时，为了得到模型的平均结果，使用十折交叉验证的方法，并去除最高和最低的一组结果，剩下的 8 个结果求平均值。

5　实验结果和评价

本文研究的 2 个问题，一是检验 *n*-gram 卷积核与基准模型和其他非卷积核模型的性能比较，二是检验融合了 Self-Attention 机制和 *n*-gram 卷积核的方法。

n-gram 卷积核对复合名词的识别（即 RQ1）的实验结果与其他模型的结果比较如表 7 所示。

表 7　*n*-gram 卷积核对模型影响的结果比较

Table 7　Comparison of the effects of *n*-gram convolution on the model

模　型		短语多词识别评价指标			短语单词识别评价指标		
		P	*R*	F_1	*P*	*R*	F_1
无卷积核	CRF (Baseline)	19.04	5.88	8.99	37.93	7.69	12.79
	Bi-LSTM+CRF	24.14	20.59	22.22	28.7	23.08	25.58
	Self-Attention+CRF	12.86	13.24	13.04	21.05	16.78	18.68
3-gram	CNN+CRF	28.00	20.59	23.73	41.46	23.78	30.22
2-3-4-gram	TextCNN+CRF	21.28	29.41	24.96	30.00	33.57	31.68

由表 7 可知：基准模型 CRF 的两项 F_1 值分别为 8.99 和 12.79。而 2 个非卷积核模型，即 Bi-LSTM+CRF，两项的 F_1 值分别为 22.22 和 25.58；Self-Attention+CRF 两项的 F_1 值分别为 13.04 和 18.68。这两种基础的神经网络模型效果均好于 CRF 基准模型。但是 3-gram CNN+CRF 的所有评价指标都显著高于基准模型和其他神经网络模型，F_1 值达到了 23.73 和 30.22。而采用三层卷积（即 2 元、3 元、4 元卷积核）的 TextCNN+CRF 模型，效果不仅好于基准模型，并且优于只有一层 3 元卷积核的 CNN+CRF 模型，F_1 值达到了 24.96 和 31.68。因此，同时使用多层 *n*-gram 卷积核，可以提升印尼语复合名词短语识别效果。

在 *n*-gram 卷积核上，融合 Self-Attention 机制（即 RQ2）的实验结果与其他模型所得的结果比较，如表 8 所示。

表 8　Self-Attention 机制和 *n*-gram 卷积核对模型影响的结果比较

Table 8　Comparison of the effects of Self-Attention mechanism and *n*-gram convolution on the model

模　型		短语多词识别评价指标			短语单词识别评价指标		
		P	*R*	F_1	*P*	*R*	F_1
2-3-gram	SHOMA(State-of-the-art)	28.12	26.47	27.27	30.77	27.97	29.30
+4-gram	TextCNN+Bi-LSTM+CRF	51.61	23.53	32.32	50.79	22.38	31.07
+4-gram+attention	TextCNN+Self-Attention+CRF	38.00	27.94	32.20	41.30	26.57	32.34

由表 8 可知：State-of-the-art 模型，即 SHOMA 模型，能够取得 27.27 和 29.30 的 F_1 值，说明其融合了 2 元、3 元卷积核的机制和序列机制，即 Bi-LSTM 的架构，在短语多词识别方面效果优于只有 *n* 元机制而缺乏序列机制的 TextCNN+CRF 模型。但 SHOMA 在短语单词识别评价指标下，并不优于无序列机制的 TextCNN+CRF。

本文提出的 2 种方法，多层 *n* 元卷积核序列机制（即 TextCNN+Bi-LSTM+CRF）的 F_1 值，比 SHOMA 有显著提高，达到了 32.32 和 31.07。而融合了 Self-Attention 机制和多层 *n* 元卷积核的方法，除了在短语多词识别上达到与 Bi-LSTM 极为接近的 F_1 值（32.20），并且在短语单词识别评价方面，取得了最高的 F_1 值为 32.34。这说明 Self-Attention 和 Bi-LSTM 虽然都可以捕获到序列信息，但 Self-Attention 在印尼语复合名词短语识别中，效率和效果都优于 Bi-LSTM；同时也说明了采用 *n*-gram 卷积核和 Self-Attention 机制对于多词表达的识别是有效的。

6　结语

本文针对印尼语复合名词短语自动识别，在统计方法和现有的 SHOMA 模型的基础上，提出了基

于多层 *n*-gram 卷积核和 Self-Attention 机制的模型（TextCNN+Self-Attention+CRF）。一系列的实验结果表明，在 CNN 层提取 *n*-gram 时，同时提取 2-garm、3-gram 和 4-gram 的特征，对印尼语复合名词短语的自动识别是有效的，能显著提升 F_1 值，取得了 32.32 和 31.07 的 F_1 值。Self-Attention 机制可以改善现有的神经网络方法，在 TextCNN+Self-Attention+CRF 模型中取得了 32.20 和 32.34 的 F_1 值，比 SHOMA 模型分别提升了 4.93% 和 3.04%。

本研究存在一些不足之处，收集数据量较少且未使用 BERT[31] 进行预训练，这对研究结果会有一定影响。在后续研究中将扩充数据量，总结印尼语复合名词的语义关系规律，并使用 BERT 对印尼语文本进行表示；进一步提升模型对印尼语复合名词短语自动识别的效果，并为其他非通用语言的多词表达识别研究提供更有价值的参考和借鉴。

参考文献：

[1] 祝慧佳 . 汉语名词复合短语识别与分类的方法研究 [D]. 哈尔滨：哈尔滨工业大学，2007.
ZHU Huijia. Research on the Methods of Chinese Noun Compounds Identification and Classification[D]. Harbin：Harbin Institute of Technology，2007.

[2] LEONARD R. The Interpretation of English Noun Phrase Sequences on the Computer[J]. Language，1987，63(2)：429.

[3] MCDONALD D B. Understanding Noun Compounds[D]. Schenley Park Pittsburgh：Carnegie Mellon University，1982.

[4] TER STAL W G. Automated Semantic Analysis of Compounds：Problems Identifications and Literature Overview[EB/OL]. [2020−01−12]. https://research.utwente.nl/en/publications/automated-semantic-analysis-of-compounds-problems-identifications.

[5] 陈昌熊 . 复合词分析及其在信息检索中的应用 [D]. 上海：上海交通大学，2008.
CHEN Changxiong. Compound Analysis and Its Application in IR[D]. Shanghai：Shanghai Jiaotong University，2008.

[6] 高华年 . 印度尼西亚语的名词结构 [J]. 暨南大学华文学院学报，2001(1)：68−76.
GAO Huanian. Nominal Structures of Indonesian[J]. Journal of College of Chinese Language and Culture of Jinan University，2001(1)：68−76.

[7] 刘鹏远，刘玉洁 . 中文基本复合名词短语语义关系体系及知识库构建 [J]. 中文信息学报，2019，33(4)：20−28.
LIU Pengyuan，LIU Yujie. Semantic Relations Hierarchy and Knowledge Base for Chinese Basic Noun Compounds[J]. Journal of Chinese Information Processing，2019，33(4)：20−28.

[8] WILLIAMS J R. Boundary-Based MWE Segmentation with Text Partitioning[EB/OL]. [2020−01−20]. https://arxiv.org/abs/1608.02025.

[9] HOSSEINI M J，SMITH N A，LEE S I. UW-CSE at SemEval-2016 Task 10：Detecting Multiword Expressions and Supersenses Using Double-Chained Conditional Random Fields[C]//Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). San Diego：Association for Computational Linguistics，2016：931−936.

[10] ROHANIAN O，TASLIMIPOOR S，KOUCHAKI S，et al. Bridging the Gap：Attending to Discontinuity in Identification of Multiword Expressions[EB/OL]. [2020−02−03]. https://arxiv.org/abs/1902.10667.

[11] TASLIMIPOOR S，ROHANIAN O. SHOMA at Parseme Shared Task on Automatic Identification of Vmwes：Neural Multiword Expression Tagging with High Generalisation[EB/OL]. [2020−01−09]. https://arxiv.org/abs/1809.03056.

[12] DOWNING P. On the Creation and Use of English Compound Nouns[J]. Language，1977，53(4)：810−842.

[13] WARREN B. Semantic Patterns of Noun-Noun Compounds[J]. Acta Universitatis Gothoburgensis，1978，41：261−266.

[14] FREDERICK J，NEWMEYER，LEVI J N. The Syntax and Semantics of Complex Nominals[J]. Language，1979，55(2)：396−407.

[15] SMADJA F. Retrieving Collocations from Text: Xtract[J]. Computational Linguistics，1993，19(1)：143−177.

[16] PIAO S，SUN G F，RAYSON P，et al. Automatic Extraction of Chinese Multiword Expressions with a Statistical Tool[C]//Workshop on Multi-Word-Expressions in a Multilingual Context in EACL 2006. Trento：[s. n.]，2006：17−24.

[17] WAKAKI H，FUJII H，SUZUKI M，et al. Abbreviation Generation for Japanese Multi-Word Expressions[C]//Proceedings of the Workshop on Multiword Expressions Identification，Interpretation，Disambiguation and Applications. Suntec：Association for Computational Linguistics，2009：63−70.

[18] 唐 亮，李 倩，许洪波，等 . 基于多策略过滤的汉日多词短语抽取和对齐 [J]. 山东大学学报（理学版），2015，50(9)：21−28.
TANG Liang，LI Qian，XU Hongbo，et al. Chinese-Japanese Multi-Word Phrase Extraction and Alignment Based on Multi-Strategy Filtering[J]. Journal of Shandong

University (Natural Science)，2015，50(9)：21−28.

[19] BALDWIN T. Deep Lexical Acquisition of Verb-Particle Constructions[J]. Computer Speech & Language，2005，19(4)：398−414.

[20] KATZ G，GIESBRECHT E. Automatic Identification of Non-Compositional Multi-Word Expressions Using Latent Semantic Analysis[C]//Proceedings of the Workshop on Multiword Expressions Identifying and Exploiting Underlying Properties. Sydney：Association for Computational Linguistics，2006：12−19.

[21] VAN DE CRUYS T，MOIRóN B V. Semantics-Based Multiword Expression Extraction[C]//Proceedings of the Workshop on a Broader Perspective on Multiword Expressions. Prague：Association for Computational Linguistics，2007：25−32.

[22] 肖　健，徐　建，徐晓兰，等 . 英中可比语料库中多词表达自动提取与对齐 [J]. 计算机工程与应用，2010，46(31)：130−134，187.
XIAO Jian，XU Jian，XU Xiaolan，et al. Automatic Extraction and Alignment of Multiword Expressions from English-Chinese Comparable Corpus[J]. Computer Engineering and Applications，2010，46(31)：130−134，187.

[23] 梁颖红，谭红叶，鲜学丰，等 . 基于改进 DE-Tri-Training 算法的汉语多词表达抽取 [J]. 数据采集与处理，2017，32(1)：141−148.
LIANG Yinghong，TAN Hongye，XIAN Xuefeng，et al. Chinese Multi-Word Expression Extraction Based Improved DE-Tri-Training Algorithm[J]. Journal of Data Acquisition and Processing，2017，32(1)：141−148.

[24] 赵维纳，李　琳，刘汇丹，等 . 藏语三音动词短语自动抽取研究 [J]. 中文信息学报，2015，29(3)：196−200.
ZHAO Weina，LI Lin，LIU Huidan，et al. Automatic Extraction of Trisyllabic Verb Phrases in Tibetan[J]. Journal of Chinese Information Processing，2015，29(3)：196−200.

[25] 麦热哈巴·艾力，阿孜古丽·夏力甫，吐尔根·依布拉音 . 维吾尔语多词表达抽取方法研究 [J]. 计算机工程与应用，2014，50(8)：26−30.
MAIREHABA Aili，AZIGULI Xialifu，TUERGEN Yibulayin. Research on Extracting Methods of Multi Word Expression in Uyghur Texts[J]. Computer Engineering and Applications，2014，50(8)：26−30.

[26] 张海军 . 维吾尔语短语自动抽取研究进展 [J]. 计算机科学与探索，2015，9(12)：1420−1429.
ZHANG Haijun. Progress of Automatic Extraction of Uyghur Phrases[J]. Journal of Frontiers of Computer Science and Technology，2015，9(12)：1420−1429.

[27] 古丽扎达·海沙，古丽拉·阿东别克 . 哈萨克语动词短语自动识别研究与实现 [J]. 计算机工程与应用，2015，51(2)：218−223，240.
GULIZADA Haisa，GULILA Altenbek. Research on Automatic Identification of Base Verb Phrases in Kazakh[J]. Computer Engineering and Applications，2015，51(2)：218−223，240.

[28] GRAVES A，SCHMIDHUBER J. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures[J]. Neural Networks，2005，18(5/6)：602−610.

[29] KIM Y. Convolutional Neural Networks for Sentence Classification[EB/OL]. [2020−02−22]. https://arxiv.org/abs/1408.5882.

[30] SAVARY A，RAMISCH C，CORDEIRO S，et al. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions[C]//Proceedings of the 13th Workshop on Multiword Expressions. Valencia：Association for Computational Linguistics，2017：31−47.

[31] DEVLIN J，CHANG M W，LEE K，et al. BERT：Pre-training of Deep Bidirectional Transformers for Language Understanding[EB/OL]. [2020−02−19]. https://arxiv.org/abs/1810.04805.

（责任编辑：邓光辉）