第49卷第7期 中国科学技术大学学报 Vol. 49, No. 7
2019年7月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jul. 2019

文章编号:0253-2778(2019)07-0517-07

马来语领域多词组无监督识别

王 琳[1],刘伍颖[2]

(1.上海外国语大学贤达经济人文学院,上海 200083;2.广东外语外贸大学语言工程与计算实验室,广东广州 510420)

摘要:多词组是一种优化的语言复用粒度.,由于一些非通用语言的多词组与词之间缺乏显式形态边界,导致多词组自动识别困难.针对马来语领域多词组识别问题,提出一种基于自然标注的无监督抽取与聚类算法.算法首先采用空格符二值分类实现变长马来语多词组抽取;然后将文档级的自然类别标注迁移到多词组级类别聚类;最后过滤掉通用多词组,萃取多个领域多词组数据集.在272 783马来语文本文档数据集上的实验结果表明,提出的算法不但能够精准地抽取多词组,而且能够高效地实现多词组领域词典聚类.

关键词:无监督识别;多词组;领域词典;自然标注;马来语

中图分类号:TP391.1 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2019.07.001

引用格式:王琳,刘伍颖. 马来语领域多词组无监督识别[J]. 中国科学技术大学学报,2019,49(7):517-523.
WANG Lin, LIU Wuying. Unsupervised identification of malay domain multiword expressions[J]. Journal of University of Science and Technology of China, 2019,49(7):517-523.

Unsupervised identification of Malay domain multiword expressions

WANG Lin[1], LIU Wuying[2]

(1. *Xianda College of Economics and Humanities, Shanghai International Studies University, Shanghai* 200083, *China*;
2. *Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou* 510420, *China*)

Abstract: Multiword expression (MWE) is an optimal granularity of language reuse. However, no explicit formal boundaries between MWEs and other words cause a serious problem on automatic identification of MWEs for some non-common languages. We address the identification issue of Malay domain MWEs, and propose a natural-annotation-based unsupervised extraction and clustering algorithm. In the novel algorithm, we firstly use a binary classification for each space character to solve length-varying Malay MWEs extraction, secondly transfer natural document-level category annotations to MWE-level ones for Malay MWEs clustering, and finally distill out several domain datasets of MWEs after filtering general MWEs. The experimental results in the Malay dataset of 272 783 text documents show that our algorithm can extract MWEs precisely and dispatch them into domain lexicons efficiently.

Key words: unsupervised identification; multiword expression; domain lexicon; natural annotation; Malay

收稿日期:2018-06-15;**修回日期**:2018-09-18

基金项目:上海市社科规划项目(2019BYY028);国家语委重点项目(ZDI135-26);广东省自然科学基金(2018A030313672);广州市人文社科重点研究基地重点项目(2017-IC-02)资助.

作者简介:王琳,女,1983年生,硕士/讲师.研究方向:计算语言学和语料库语言学. E-amil: lwang@xdsisu.edu.cn

通讯作者:刘伍颖,博士/副教授,硕士生导师,E-amil: wyliu@gdufs.edu.cn

0 引言

对于机器翻译及其他自然语言处理算法，语言复用和处理粒度一直令人困惑[1]. 多词组(multiword expression, MWE)，也称多词表达，是指由两个以上词汇串联成的语言片段，其形态粒度介于词汇与句子之间，其语义相对固定并超越每个构成词汇的意义，是一种更加高效的语言复用和处理粒度[2].

目前，针对英语、汉语等通用语言的多词组识别研究比较丰富[3-4]，产生了许多基于规则和基于统计的识别算法[5-6]. 近来针对非通用语言的多词组识别研究兴起[7]，逐渐形成当前极具挑战的研究热点. 现代马来语(Bahasa Melayu)属于非通用语言，其字母表采用类似英语的标准拉丁字母表. 目前全球约有2.9亿人使用马来语，主要分布于文莱、印度尼西亚、马来西亚和新加坡. 随着语言科学和计算技术的飞速发展，互联网上马来语文本文档数量也快速增长.

互联网上带有自然标注的语言大数据给非通用语言多词组抽取与聚类等带来新的重大机遇. 本文针对马来语领域多词组识别问题，提出一种基于自然标注的无监督抽取与聚类算法，实现马来语多词组领域词典自动构建.

1 相关研究

多词组识别的研究最早主要是针对通用语言开展的. 例如，有监督(supervised)算法采用统计上下文特征在通用语言上能够达到较高的识别准确率[8]；半监督(semi-supervised)算法采用句法特征构建阿拉伯语口语多词组资源[9]. 这些有监督和半监督算法通常需要人工标注，标注过程费时费力，难以跟上大数据时代语言信息的增长速度.

近来在研究越南语多音节词识别时发现，多音节词识别与多词组识别面临的科学问题非常相似，无监督算法也是有效的[10]. 无监督算法是一种应对语言大数据真实场景的简洁高效模型. 其有效性归因于大数据时代，简单模型加大量数据优于复杂精细模型加少量数据[11]. 无监督学习不需要人工标注[12]，更加适合从动态语言大数据中抽取多词组.

目前，有研究采用互信息和信息熵从文本文档中识别多词组，并取得了较好的效果[13-14]，该思想对领域多词组识别具有重要的参考价值. 例如，国内民族语领域多词组自动抽取技术[15]；将非通用语言的语汇化问题转化为空格符二值分类问题[16]；采用多实例学习将粗粒度文档级标注迁移到细粒度实例级标注[17]. 该思想对领域多词组识别具有重要的参考价值。.

受上述研究的启发，本文设计了一种新的无监督架构. 一方面充分发挥语言大数据加简洁高效模型的优势，采用空格符二值分类算法实现多词组抽取；另一方面充分利用文档级自然标注实现多词组领域聚类.

2 无监督架构

本文提出的用于马来语领域多词组识别的无监督架构，主要包括抽取(extracting)和聚类(clustering)两部分，如图1所示. 抽取部分接收大规模马来语文本文档，生成马来语多词组. 聚类部分接收抽取部分生成的马来语多词组，输出多个领域的多词组词典以及通用多词组词典. 这些领域词典资源和通用词典资源可用于各种自然语言处理应用. 该架构是一种灵活的元结构，各类具体的高效抽取与聚类算法都可以集成.

现代马来语文本中的空格符与其他拉丁字母语言中的空格符在功能上具有相似性，都是一种重载符：既可以充当多词组内部的连接符，又可以充当多词组之间的分隔符. 基于这一思路，在架构抽取部分，将马来语多词组抽取定义为任一空格符的二值分类任务，并通过5个处理单元共同完成抽取. 马来语多词组只会由马来语单词串联而成，可能包含马来语多词组的文本也只会是纯马来语词串片段. 首先，当马来语文本文档到达，触发片段预处理(fragment preprocessing)单元，预处理将马来语文本分割成纯马来语词串序列，每个词串片段内部不包含标点、拉丁语单词以及其他符号. 然后，词频统计(word frequency counting)单元和词汇级Bigram频率统计(Bigram frequency counting)单元分别接收生成的片段，并行统计相应的频率；空格二值分类(space binary classifying)单元接收片段以及词频、Bigram频率统计结果，利用本文提出的空格二值分类法对每个空格进行分类，输出带标注的片段；当空格是多词组内部的连接符时，在输出标注片段中用连接线替换该空格；当空格是多词组之间的分隔符时，在输出标注片段中维持原空格. 最后，多词组收

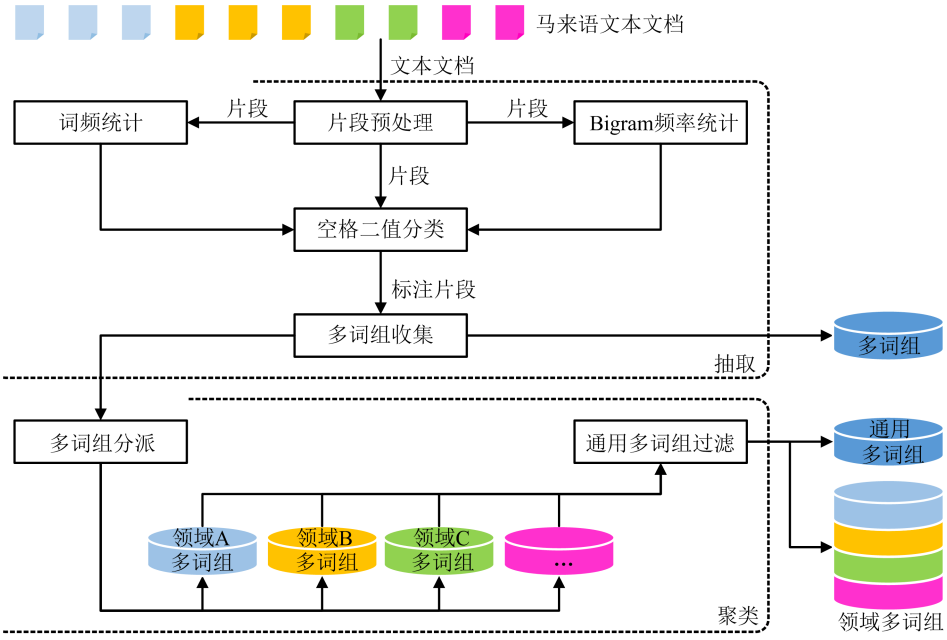

图 1　无监督抽取与聚类架构

Fig. 1　Unsupervised extracting and clustering architecture

集(multiword expression collecting)单元根据空格切分标注片段,并对至少包含一个短连线的多词组进行频率统计,再根据预设的频率阈值(本文中该阈值为 3)输出高频多词组.

二值分类方法的有效性很大程度上取决于马来语文本文档的数量. 随着语言大数据的兴起,互联网不断涌现出大量的马来语新闻网页,通过自动抓取和清洗过滤能够获得大规模马来语文本数据集. 不仅如此,大部分新闻网页还自带人工编辑的类别标注元信息. 这种自然的文档级类别标注可以下沉迁移到多词组级类别聚类,形成多词组领域. 架构聚类部分只包含 2 个处理单元. 多词组分派(multiword expression dispatching)单元根据接收的多词组源网页类别标注命名增量创建的多词组领域,并将多词组分派到相应的领域. 通用多词组过滤(general multiword expression filtering)单元将领域覆盖较高的多词组分离出来形成通用多词组集合,再从每个领域多词组集合中过滤掉通用多词组,生成多个领域多词组集合.

3　算法

针对马来语领域多词组识别问题,我们基于上述无监督抽取与聚类架构,设计了马来语多词组抽取与聚类算法,如算法 3.1 所示.

算法 3.1　马来语多词组抽取与聚类算法

```
1 //马来语多词组抽取与聚类算法
2 input: document[] mtds; //马来语文本文档
3     float ct; //粘度阈值
4 output: cluster[] dmwes; //领域多词组
5 function string[]: extracting(document[] mtds; float ct)
6 string[] frags;
7 map<string, integer> bf;
8 map<string, integer>wf;
9 for integer i←1 to mtds.size do
10    string[]frag←fragment preprocessing(mtds[i].
      text); //片段预处理
11    frags.merge(frag);
12    bf.merge(bigram frequency counting(frag)); //
      bigram 频率统计
13    wf.merge(word frequency counting(frag)); //
      词频统计
14 end for
15 //空格二值分类
16 for integer j←1 to bf.key set.size do
17    integer bwf←bf.get(bf.key set[j]);
18    integer fwf←wf.get(bf.key set[j].first word);
19    integer swf←wf.get(bf.key set[j].second word);
20    float c←(bwf/fwf+bwf/swf)/2;
21    if(c>ct)
22    then frags.update(bf.key set[j].first word+'-
      '+bf.key set[j].second word);
23    end if
24 end for
25 string[] mwes←multiword expression collecting
```

```
    (frags); //多词组收集
26 return mwes.
27
28 function cluster[]: clustering(document[] mtds;
   string[] mwes)
29 //多词组分派
30 for integeri←1 to mwes. size do
31     document[] do cs←mtds. get documents(mwes
       [i]);
32     for integer j←1 to do cs. size do
33       string category←do cs[j]. category;
34       if (dmwes. contain(category))
35       then dmwes. get(category). add(mwes[i]);
36        else dmwes. put(cluster. new(category). add
         (mwes[i]));
37       end if
38     end for
39 end for
40 //通用多词组过滤
41 string[] gmwes←collecting general(dmwes); //通
   用多词组
42 for integer i←1 to dmwes. size do
43     dmwes[i]. remove(gmwes);
44 end for
45 return dmwes.
```

算法 3.1 伪码中的 extracting 函数与 clustering 函数分别对应于图 1 架构中的抽取与聚类两个部分. 该算法的主要创新在于粘度计算的思想. 本文认为两个邻接单词之间空格的粘度正比于两个单词邻接共现概率,反比于空格前后的单词单独出现概率,因此可以通过两个单词邻接共现概率除以单词单独出现概率来计算. 具体实现时,本文等权重对待空格前后的单词,设计出算法 3.1 第 20 行的粘度计算公式. 这一思想理论上可以实现任意词数的多词组抽取,且只需通过词频和 Bigram 频率统计就能较好解决不定长多词组抽取问题. 该算法还充分利用文档级自然标注实现多词组聚类. 在通用多词组过滤过程中,本文将领域覆盖阈值设为 0.6,即 60%以上领域多词组是通用多词组. 领域覆盖阈值设为 0.6 是采用梯度阈值设置实验加上语言专家人工标注得出的最优阈值,大于或小于 0.6 都会降低领域多词组的有效性.

整个算法执行过程的空间开销主要由单词-频率索引和 Bigram-频率索引决定,不会随文本文档数量增加而线性增长,目前的存储器完全可以接受. 由于本文算法是流处理算法,对每个文档仅需扫描一遍增量构建上述两个索引,因此算法的时间开销正比于文本文档数,在时间上也是高效的. 综上可知,马来语多词组抽取与聚类算法的时空复杂度在实际的马来语领域多词组无监督识别应用中是可接受的.

4 实验

为了验证本文无监督抽取与聚类架构和算法的有效性,首先抓取互联网网页新闻,制备大规模马来语文本文档集合. 实验总共抓取了 272 783 篇带类别标注的马来语新闻网页. 然后利用已有的 85 539 条马来语多词组标准答案对实验结果进行评价. 为了进一步验证本文方法的优越性,同时运行当前主流的基于互信息和信息熵的基线算法[14]. 最后得出实验结果并进行比较分析.

4.1 多词组抽取结果讨论

在多词组无监督抽取实验中,我们从 0.1 到 0.9 梯度设置粘度阈值,接着报告算法每次运行结果的准确率(P)、召回率(R)、F_1 值(F_1),以此评价马来语多词组抽取与聚类算法的抽取效果. 图 2 描绘了上述 3 项指标的详细变化趋势. 由图 2 可知,随着粘度阈值从小变大,P 值和 R 值呈相反趋势变化,准确率逐渐提高,召回率不断降低;当粘度阈值等于 0.3 时,F_1 值达到最高峰值,其中 P、R 和 F_1 值分别为 0.296 5、0.231 1 和 0.259 7. 在完全相同的环境下,基线算法最优运行结果 P、R 和 F_1 值分别为 0.069 3、0.301 3 和 0.112 6. 实验结果说明,本文无监督抽取算法的效果优于基于互信息和信息熵算法的效果.

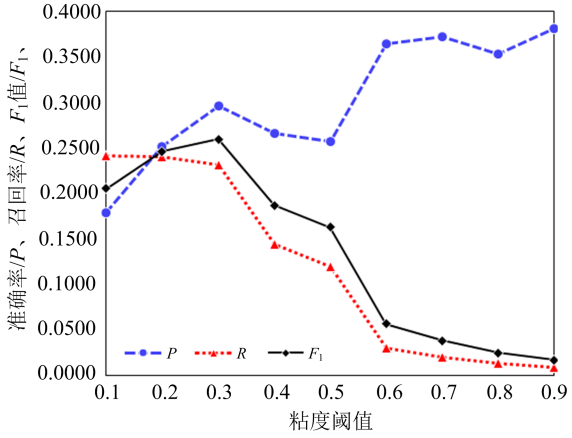

图 2 抽取的评价结果

Fig. 2 Evaluation result of extraction

图 2 表明,在最优 F_1 值情况下,本文算法的准确

率只有 30%左右，其中一个主要原因是马来语文本训练语料与标准答案之间是相互独立的，而实际抽取准确率应该高于实验结果；当 F_1 值等于最优的 0.259 7 时，本文算法共抽取出 66 625 条马来语多词组，我们从中简单随机采样出的 100 条多词组样例，如表 1 所示.

表 1　马来语多词组部分样例

Tab. 1　Partial examples of Malay MWEs

abdulhalim	daniilkvyat	kacang soya	montkiara	sanisahhuri
adriansutil	desperate housewives	kampung gesirtengah	nacho monreal	selenagomez
ajayjayaram	dzulkeflyahmad	kaspersky lab	nicklasbendtner	shahurainabusamah
alatpenimbang	emmanueladebayor	keranamumalaysia	nursultannazarbayev	silatcekak
alexkomang	faezahelai	ketuananmelayu	orang bunian	skalarichter
amitabhbachchan	forrest gump	kohilanpillay	padangmahsyar	starhill gallery
angkasalepas	gagahberani	kotamarudu	papantanda	sungaibedaun
asafapowell	gas pemedihmata	kubangkerian	pasukanpetugas	sungaibesar
azlanmisron	giorgionapolitano	laurentciman	pengirananaksaleha	surah faatir
bahasapengantar	gununglawu	lemparcakera	penglipurlara	tanjungbungah
bandarayakuching	hailegebrselassie	liewdaren	persipurajayapura	tapaksemaian
benjaminmoukandjo	hartasepencarian	louis van gaal	polis dirajamalaysia	telukintan
berterusterang	holgerbadstuber	lucasdigne	profesormadya	tikambatu
bolakeranjang	igorakinfeev	malalayousafzai	pulaukerengga	tudung saji
bolasepak	ikanbilis	marcostorari	qiao bin	tunjukperasaan
bukitgoram	istananuruliman	masjid rahmaniah	raisyatim	universitithammasat
cacar air	jalurgemilang	mehmetdurakovic	rekabentuk	volkandemirel
chanpeng soon	jemaahislamiyah	mengambilalih	rolandgarros	wong mew choo
chenderongbalai	john obi mikel	milakunis	ryan babel	yongkhoonseng
cristianoronaldo	junior hoilett	misanoadriatico	sanaksaudara	zhangjike

分析表 1 中的样例可知，抽取出的多词组类别既包括人名、地名、机构名等专有概念多词组，也包括普通概念多词组. 由于实验中的片段预处理只对英语单词进行了切片触发，所以其他拉丁语单词或拉丁语拼音都得到保留，使得 qiao bin(乔斌)、wong mew choo(黄妙珠)、yongkhoonseng(杨昆贤)、zhangjike(张继科)等人名也得以识别. 尽管这些人名与严格意义上的马来语多词组不完全相符，但能够正确识别人名对马来语处理而言还是具有积极意义的. 语言专家对实验结果进行了人工标注，如果把马来语外来词组一并统计，本文抽取算法的实际准确率高于 80%.

4.2　多词组聚类结果讨论

在多词组无监督聚类实验中，我们根据多词组源网页类别标注聚类上述 66 625 条马来语多词组，最终聚类出 9 个领域马来语多词组词典以及 1 个包含 3 419 条马来语多词组的通用词典. 具体 10 个词典中的马来语多词组数量如表 2 所示.

表 2　马来语多词组数量

Tab. 2　Number of Malay MWEs

通用多词组数量	领域多词组	
	领域标注	数量
3 419	Semasa (时事)	10 797
	Dunia (国际)	8 284
	Nasional(国内)	16 023
	Sukan(体育)	13 531
	Jenayah (犯罪)	5 733
	Utusan Borneo-Berita Nasional (婆罗洲新闻. 国内)	12 739
	Utusan Borneo-Berita Sarawak (婆罗洲新闻. 沙捞越)	20 530
	Utusan Borneo-Berita Iban(婆罗洲新闻. 伊班族)	14 906
	Utusan Borneo-Sukan (婆罗洲体育)	7 628

为了更加直观地分析领域马来语多词组的聚类效果,本文采用简单随机采样方法,从聚类出的13 531条体育领域多词组中抽出表3所示的100条多词组样例.分析表3可知,体育领域多词组相对收敛且领域相关.例如,lompat tinggi(跳高)、rejam lembing(标枪)是体育运动名;gamba osaka(大阪)、persija jakarta(雅加达)是足球俱乐部名;jia yifan(贾一凡)和low wee wern(刘薇雯)分别是羽毛球运动员和壁球运动员等.尽管表3中人名占据较大比例,但这些人主要是运动员和体育相关人员,领域聚类效果明显.

表3 体育领域多词组部分样例

Tab.3 Partial examples of sukandomain MWEs

abdulhamid	daphne iking	kasperschmeichel	minatsumitani	sebastianvettel
afrikaselatan	debbysusanto	kenichitago	mohamedelneny	sergiroberto
airasia x	diafrasakho	kentomomota	muhammadubuhari	srikanthkidambi
amerikasyarikat	didierdeschamps	lassanadiarra	nicorosberg	stevemandanda
andriypyatov	edwinekiring	lewisholtby	nigelmansell	stevennaismith
arkadiuszmilik	ennervalencia	lompattinggi	nityakrishindamaheswari	sunway lagoon
ayakatakahashi	gambaosaka	low wee wern	novipazar	susurgalur
azzuddinahmad	gerardomartino	lucasdigne	penyakitparkinson	theowalcott
bertegursapa	gunnarnordahl	luissuarez	persijajakarta	thiagomotta
bixentelizarazu	hassanchaito	lukaspodolski	portoalegre	tian houwei
brendanrodgers	iagoaspas	mahalijasuli	rabiatuladawiyah	tonikroos
briceleverdez	jerzydudek	mannypacquiao	rachaelgrinham	tulangselangka
brown ideye	jiayifan	marc wilmots	radio televisyenbrunei	vincentkompany
bryannicksonlomas	johancruyff	marcomelandri	radjanainggolan	vitomannone
bukit jambul	joleonlescott	marcostreller	rayovallecano	wilfriedzaha
calum chambers	jonnyevans	marcosrojo	razipismail	yasuhito endo
chanpeng soon	julianbaumgartlinger	mariobalotelli	rejamlembing	yelenaisinbayeva
chidiedeh	julung kali	marouanefellaini	roberto soldado	zhang nan
chloemagee	kamarudinmeranun	mathewleckie	ryanshotton	zinedinezidane
clintdempsey	kamilgrosicki	mathieuflamini	sameerverma	zulfahmikhairuddin

综合上述实验结果可以看出,本文的主要贡献:一是成功地将马来语多词组抽取问题转换成空格符二值分类问题,并通过简洁的粘度计算有效地实现了不定长多词组抽取;二是创新马来语领域多词组无监督聚类思想,成功地将文档级的自然类别标注高效迁移到多词组级类别聚类,并且支持领域多词组资源的增量构建.

5 结论

本文提出了一种基于自然标注的多词组抽取与聚类算法,并成功地实现了马来语领域多词组资源构建.实验结果表明,本文算法的效率仅取决于单词共现频率这种简洁的上下文知识形态和自然的文档级类别标注.网络语言大数据产生了大量带自然标注的文本文档,使得这种无监督算法拥有更加广阔的应用前景.进一步研究将关注附加语言知识对空格二值分类效果的影响,如拉丁字母停用词、形态、句法、语义等知识,我们准备将多词组抽取方法移植应用于印度尼西亚语、菲律宾语等其他适合的南岛语系.此外,领域多词组无监督聚类思想可以进一步深化,我们准备尝试先对文本文档进行多层次聚类,实现不同粒度的结构化领域多词组资源构建.

参考文献(References)

[1] AITI A, ALJUNIED S M, LI H Z. Malay multi-word expression translation [C]// Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation. Singapore, ACM, 2009: 21-24.

[2] DE CASELI H M, RAMISCH C, DAS GRAÇAS VOLPE NUNES M, et al. Alignment-based extraction of multiword expressions[J]. Language Resources and Evaluation, 2010, 44(1):59-77.

[3] 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.

[4] 龚双双，陈钰枫，徐金安. 基于网络文本的汉语多词表达抽取方法[J]. 山东大学学报，2018，53(9)：40-48.

[5] DUBREMETZ M, NIVRE J. Extraction of nominal multiword expressions in French[C]// Proceedings of the 10th Workshop on Multiword Expressions. Gothenburg, Sweden: IEEE, 2014: 72-76.

[6] EBRAHIM SHEGAZY D, GADAL-HAQQ M M, et al. Detecting and integrating multiword expression into English-Arabic statistical machine translation [J]. Procedia Computer Science, 2017, 117:111-118.

[7] KUMAR S, BEHERA P, JHA G N. A classification-based approach to the identification of multiword expressions (MWEs) in Magahi applying SVM[J]. Procedia Computer Science, 2017, 112:594-60.

[8] FARAHMAND M, MARTINS R. A supervised model for extraction of multiword expressions based on statistical context features [C]//Proceedings of the 10th Workshop on Multiword Expressions. Gothenburg, Sweden: IEEE, 2014: 10-16.

[9] AL-BADRASHINY M, HAWWARI A, GHONEIM M, et al. SAMER: A semi-automatically created lexical resource for arabic verbal multiword expressions tokens paradigm and their morphosyntactic features [C]// Proceedings of the 12th Workshop on Asian Language Resources. Osaka, Japan: IEEE, 2016: 113-122.

[10] LIU Wuying, WANG Lin. Vietnamese multisyllabic-word extraction for word segmentation [J]. International Journal of Asian Language Processing, 2017, 27(1):61-77.

[11] HALEVY A, NORVIG P, PEREIRA F. The unreasonable effectiveness of data[J]. IEEE Intelligent Systems,2009, 24(2):8-12.

[12] VLACHOS A. Evaluating unsupervised learning for natural language processing tasks[C]// Proceedings of the First Workshop on Unsupervised Learning in NLP. Edinburgh, UK: IEEE, 2011: 35-42.

[13] 李丽双，王意文，黄德根. 基于信息熵和词频分布变化的术语抽取研究[J]. 中文信息学报，2015，29(1)：82-87.

[14] 刘剑，唐慧丰，刘伍颖. 一种基于统计技术的中文术语抽取方法[J]. 中国科技术语，2014，16(5):10-14.

[15] 田生伟，钟军，禹龙. 维吾尔语多词领域术语的自动抽取[J]. 中文信息学报，2015，29(2):133-141.

[16] LIU Wuying. Supervised ensemble learning for Vietnamese tokenization[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2017, 25(2):285-299.

[17] ZHANG Yi, SURENDRANA C, PLATTJ C, et al. Learning from multi-topic web documents for contextual advertisement[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA: Springer, 2008: 1051-1059.