Skip-Gram 模型融合词向量投影的微博新词发现①

于　洁

(福建信息职业技术学院 计算机工程系, 福州 350003)

摘　要: 随着微博等社交网络的普及, 新词源源不断涌现, 分词系统经常将新词错误切分为单字. 新词发现已经成为中文自然语言处理领域的研究热点. 现有新词识别方法依赖大规模语料统计数据, 对低频新词识别能力差. 本文提出一种扩展Skip-gram模型和词向量投影方法, 将两者结合后能缓解自然语言处理中常见的数据稀疏问题, 有效识别低频新词,进而提高分词系统的准确率和召回率.

关键词: skip-gram; SOM; 词向量; 微博; 新词发现

Microblog New Word Recognition Combining Skip-Gram Model and Word Vector Projection

YU Jie

(Computer Engineering Department, Fujian Polytechnic of Information Technology, Fuzhou 350003, China)

Abstract: With the popularity of microblog and other social networks, a steady stream of new words emerge, Chinese word segmentation systems often cut the new words into Chinese characters. The new word discovery has become a hot topic in the field of Chinese natural language processing. Existing new word recognition methods rely on the statistical data of large-scale corpus, the ability of new low-frequency word recognition is poor. This paper presents an extension of skip-gram model and word vector projection method, after the combination of the this two methods can ease the data sparseness problem effectively in natural language processing, to identify new low-frequency words, and to improve the precision and recall rate of Chinese word segmentation system.

Key words: skip-gram; SOM; word vector; microblog; new word recognition

随着互联网的高速发展和社交网络的迅速普及, 微博、微信等新兴社交媒体的出现代表着自媒体时代的到来. 根据新浪公布的数据, 微博注册用户已超 5 亿, 2015 年 3 月, 微博月度活跃用户数量为 1.98 亿, 较上年同期增长 38%. 在每天产生的上亿条微博中, 网民大众创造的新词源源不断地涌现, 如“新常态”、“沪港通”、“熊孩子”、“奇葩”、“占中”、“抗埃”等等. 这些新词经常用于表达各种新生事物、丰富情感色彩和突发事件话题.

中文自动分词是中文自然语言处理领域众多研究的基础性工作. 现有分词系统往往将未登录词错误切分成单字或已登录词. 据统计, 未登录词造成的分词精度失落比分词歧义大 5～10 倍以上[1].

近年来, 移动社交网络中频繁出现的新词不仅对现有自动分词技术带来新的挑战, 也对微博用户情感分析、网络舆情热点发现等社交网络挖掘产生重要影响. 中文新词发现研究日益受到重视.

1　相关研究

新词是指随社会发展而新出现的词语或原有词语被赋予新的含义. 未登录词(Out-of-vocabulary, 简称OOV)是指在当前词典中没有出现的词. 新词亦属于未登录词, 一般研究对此不作区分[2].

大部分新词发现研究主要是从语言规范性相对较高的传统语料中发现新词, 但是上亿微博发布者的知识背景和用词习惯参差不齐, 且受单条微博字数限制

① 基金项目:福建省教育厅科技项目(JA11304)

收稿时间:2015-11-17;收到修改稿时间:2015-12-21　[doi:10.15888/j.cnki.csa.005236]

而倾向口语化表达. 因此, 在微博等社交网络上的新词发现比在一般语料上更加困难.

现有研究可以分为两大类, 基于规则的方法和基于统计的方法.

基于规则的方法主要根据中文构词原理制作规则库, 对新词进行匹配. 早期郑家恒[3]等人以现代汉语构词法为原则, 以现代汉语构词规范为标准, 建立了构词规则库对新词语进行识别. 崔世起[4]等人针对新词常见模式用垃圾词典和词缀词典对候选新词进行垃圾过滤, 最后使用词性过滤规则等进一步过滤, 基准实验的 F-Measure 达到 82.22%. 基于规则的新词发现方法准确率、召回率虽然较高, 但随着近年社交网络新词频出, 人工建立规则成本高、效率低, 且规则间容易出现冲突.

基于统计的方法主要通过统计语料中频度信息, 如互信息, 邻接类别, 邻接熵等来识别新词. 罗盛芬[5]等人考察了九种常用统计量在汉语自动抽词中的 表现, 建议直接选用互信息进行自动抽词. Feng[6]等人提出了一个衡量字符串上下文的自由程度的统计量邻接类别(Accessor Variety)进行新词识别. Li[7]、Goh[8]等人采用一个或多个分类器识别新词, 多重分类器的 F-measure 可以达到 82%. Ye[9]等人提出了一种新的统计量重叠类别(Overlap Variety)进行新词识别.

为了提高新词识别的效果, 研究者倾向将基于规则和基于统计的方法相结合. 霍帅[10]等人获取新浪微博作为研究语料, 提出引入词关联性信息的迭代上下文熵算法, 通过上下文关系获取新词候选列表进行过滤, 并引入词法特征, 提出与统计特征相结合的过滤方法, F-Measure 达到 89.6%. 廖健[11]等人构造了 8 种构词规则, 并使用互信息统计量, 采用逐步迭代的方法识别新词, 在COAE2014 提供的 1000 万条大规模微博新词标注评测中以 F-Measure 达到 20.7%取得第一名的成绩.

2 语言模型

语言模型是自然语言处理的基础, 常见的语言模型包括 n-gram、最大熵模型和近年深度学习研究中的神经网络语言模型等.

2.1 Skip-gram 语言模型

n-gram 模型是基于 n-1 阶马尔科夫假设的, 在自然语言处理、情感分析、文本分类、聚类等领域中应用广泛且行之有效, 但是存在数据稀疏的缺陷[12].

David Guthrie 等学者提出了改进的 skip-gram 模型[13], 不仅允许一个词与相邻词组合, 还允许与附近 k 个词跳跃组合, 从而降低数据稀疏的影响.

假设一个句子 S 由 T 个词 $w_1,w_2,...,w_T$ 组成, 则可定义 k-skip-n-gram 为:

$$K=\left\{w_{t_1},w_{t_2},...,w_{t_n}\mid\sum_{i=1}^{n-1}t_{i+1}-t_i\le k+1\right\} \tag{1}$$

当 k=0 时, k-skip-n-gram 等同于 n-gram 集合. 通常在 n-gram 模型应用中, trigram(n=3)常用于百万词的较大规模语料, bigram(n=2)则适用小规模语料.

Skip-gram 模型的目标是让句子中的一个词能预测周围的词[14], 当 n=2 时目标函数是最大化平均对数概率:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-k-1\le j\le k+1,\,j\ne 0}\log p\left(w_{t+j}\mid w_t\right) \tag{2}$$

2.2 词向量

采用分布式表示(Distributed Representation)的方法将文本中的词转换为词向量, 能有效克服自然语言处理中的数据稀疏问题和维数灾难问题.

Bengio 等最早提出了神经网络概率语言模型[15](Neural Probabilistic Language Model). 设子序列 $w_1^{t-1}=(w_1,w_2,...,w_{t-1})$, 目标是学习一个良好的模型:

$$f\left(w_t,w_{t-1},...,w_{t-n}\right)=\hat{P}\left(w_t\mid w_1^{t-1}\right) \tag{3}$$

Google 的 Mikolov 等进一步改进, 采用 Log-linear 结构, 提出了 CBOW(Continuous Bag-of-Words)模型[14]. 该模型简化了神经网络语言模型中最耗时的非线性隐藏层, 所有词共享隐藏层. CBOW 模型与 Skip-gram 模型正好相反, 采用当前词 w_t 的上下文来预测 $p\left(w_t\mid w_{t-k},w_{t-(k-1)}...,w_{t-1},w_{t+1},w_{t+2},...w_{t+k}\right)$.

2.3 Super-organizing Maps

Kohonen[16]根据大脑皮层工作原理最早提出了自组织神经网络模型(Self-Organising Maps, SOM). SOM 是一种无监督训练神经网络, 可以将高维特征投影到一维或二维, 并能使相似的特征被映射到领近区域, 保持原有特征拓扑结构不变.

Ron[17]等人对经典的 SOM 进行了扩展, 提出了 Super-organizing Maps. 当训练数据具有多种类型特征的时候, 允许将不同类型的特征分别映射到不同的 SOM 层, 最后再按照公式(4)来确定获胜神经元.

$$D(o,u)=\sum_i\alpha_iD_i(o,u) \tag{4}$$

其中 $D(o,u)$ 表示训练对象 o 到神经元 u 的距离，α_i 为第 i 层 SOM 的权重.

3 新词发现方法

有微博例句: “(4)月宫一号志愿者 105 天实验后出舱, 氧气、水、食物实现自循环. ”, 采用中国科学院研发的汉语词法分析系统 ICTCLAS 2015 版对句子分词的结果是: “(/wkz 4/m)/wky 月宫/n 一/m 号/q 志愿者/n 105/m 天/qt 实验/n 后/f 出/vf 舱/n , /wd 氧气/n 、/wn 水/n 、/wn 食物/n 实现/v 自/p 循环/vn . /wj”.

该微博在本文实验环境中包含的 3 个未登录词(月宫一号、出舱、自循环)均被错误切分.

以新词识别常用的互信息[5,11]统计量为例, 相邻两个词 w_iw_{i+1} 的互信息关联度计算公式如下:

$$A(w_i,w_{i+1})=\log\frac{P(w_iw_{i+1})}{P(w_i)\cdot P(w_{i+1})},i=1,2,...,n-1 \qquad (5)$$

其中 $P(w_i)$ 和 $P(w_iw_{i+1})$ 分别表示 w_i 和 w_iw_{i+1} 在语料中出现的概率. 若 A(自, 循环) 大于某个经验值时, 则可以认为“自循环”是一个新词.

如果在整个语料大多数新词仅出现 1 次, 新词对应的 $P(w_iw_{i+1})$ 的值很低, w_i 与 w_{i+1} 的相邻被认为接近于偶然, 上述统计方法将失效.

通过观察已分词语料, 我们可以发现:

① 新词左右一般伴随着已登录词;

② 新词被正确切分时, 新词和上下文保持正常的语言习惯, 如“实现自循环”使用了“做|什么”的习惯用语, 而新词被错误切分时, 产生的单字或词与上下文的衔接明显不符合正常语序.

③ 新词和已登录词都符合一定的成词规则, 如“自循环”和“自启动”都使用了同一种构词规则.

本文尝试对中小规模微博语料中的低频新词进行发现, 从统计角度判断候选词和周边词是否符合正常用语模式, 并结合构词规则来识别新词.

3.1 词向量特征投影

词向量能表达一定的语义信息和语法信息[14]. 本文提出以 n-gram($n\in\{1,2\}$)的形式将词向量特征送入自组织神经网络进行无监督训练, 在降维的同时使语义相近或语言模式相近的 n-gram 投影到相近神经元, 见图 1.

将词到词向量特征的转换定义为函数 $V(\cdot)$, 根据相同的原理亦可获得词性向量特征, 词向量和词性向量按照一定权值分别投影到 Super-organizing Maps 的不同 SOM 层中. n=1 时, 词 w_t 转换为词向量特征 $V(w_t)$. n=2 时, w_tw_{t+1} 的词向量特征由 $V(w_t)$ 和 $V(w_{t+1})$ 连接而成.

新词的词向量采用已登录词的词向量加和来近似表达.

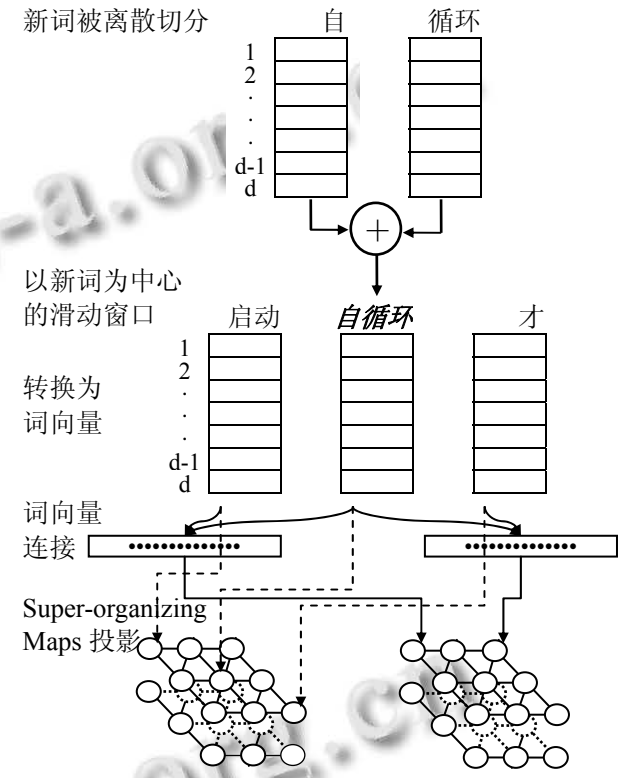

图 1　词向量特征投影(1-gram、2-gram)

定义函数 $S(\cdot)$, 表示将词转换为词向量特征再投影到 SOM 中, 获得相应神经元编号. 在本文实验中, 将训练语料以 2-gram 词向量形式投影到两层 Super-organizing Maps(20×20 神经元)中, 使具有类似用语模式的 2-gram 聚集在相同或邻近神经元上, 部分投影效果如表 1 所示.

表 1　词向量特征投影 Super-organizing Maps

1-gram SOM		2-gram SOM		
w_t	$S(w_t)$	w_t	w_{t+1}	$S(w_tw_{t+1})$
月		甲板	上	
探月		井	里	
日月塔	37	母乳	内	2
月牙		眉	间	
月柿		年	中	

出			历年	同期	
出笼			十一	期间	
出庭	346		去年	夏季	6
漏出			20 时	前后	
出站			现在	为止	
自			为	中国	
自检			到	奥地利	
自毁	117		组织	乌克兰	57
源自			通过	长江	
自验			经过	长白山	
1			信息	公开	
一			系统	启动	
2	76		黑客	破解	63
1—7			技术	咨询	
1. 83%			网络	约车	
宫殿			驾驶	很	
仓			世上	最	
坝	230		底线	已	380
框框			幼崽	却	
庭			言论	只	

3.2 扩展 Skip-gram

一个词序列由 $w_1,w_2,...,w_T$ 组成, 定义两两相邻的词组合为扩展词 w_tw_{t+1}, 则可以得到扩展词序列 $w_1w_2,w_2w_3,...,w_{T-1}w_T$. 假设长度为 $T(T>=5)$的词序列中仅有一个新词 w_{new}.

定义扩展 k-skip-n-gram 为集合:

$$K\bigcup\left\{w_{t_1}w_{t_2},...,w_{t_{n-1}}w_{t_n}\mid\sum_{i=1}^{n-1}t_{i+1}-t_i=1\right\} \tag{6}$$

以分词序列“食物|实现|自循环| . | SEND”为例, 扩展 1-skp-2-gram={“食物|实现”, “实现|自循环”, “自循环|. ”, “. |SEND”, “食物|自循环”, “实现|. ”, “自循环|SEND”, “食物实现|实现自循环”, “实现自循环|自循环. ”, “自循环. |. SEND”}.

k=1 时, 模型目标是最大化新词所在的投影序列的平均对数概率:

$$SOMLogP(w_1,w_2...,w_T)=\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-1\le j\le 1,j\ne 0\\0\le i\le 2,i\ne 1}}\begin{pmatrix}\alpha\log p\left(S(w_{t+j})\mid S(w_t)\right)+\\ \alpha\log p\left(S(w_t)\mid S(w_{t+j})\right)+\\ \beta\log p\left(S(w_{t-j})\mid S(w_{t+j})\right)+\\ \log p\left(S(w_{t-1+j+i}w_{t+j+i})\mid S(w_{t-1+i}w_{t+i})\right)+\\ \log p\left(S(w_{t-1+i}w_{j+i})\mid S(w_{t-1+j+i}w_{t+j+i})\right)+\\ \log p\left(S(w_{t-j+i}w_{t+1-j+i})\mid S(w_{t-1+j+i}w_{t+j+i})\right)\end{pmatrix} \tag{7}$$

由于新词的词向量特征并不准确, 引入参数 α 缩小新词与相邻词之间的对数概率, 参数 β 放大新词左右词之间的对数概率, 如图 2.

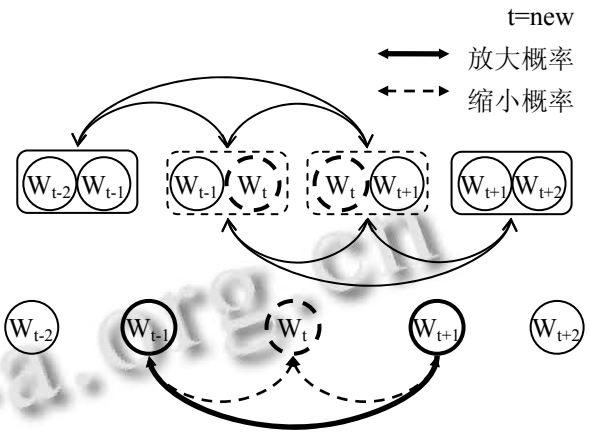

图 2 新词的对数概率及其缩放

扩展 Skip-gram 模型结合了 Skip-gram 和 CBOW 的思想, 并允许相邻的两个词合并为扩展词, 通过无监督的SOM投影, 在降低数据稀疏性的同时获取普遍性的语言使用模式. 模型假设词序列(5 个词)中只有一个新词, 适用于微博短文本的新词发现.

以微博“(4)月宫一号志愿者 105 天实验后出舱, 氧气、水、食物实现自循环. ”为例, 在本文实验环境中包含 3 个未登录词(月宫一号、出舱、自循环), 表 2 列出了标准切分词序列和 ICTCLAS 系统切分词序列的目标函数值, 标准切分的 *SOMLogP* 值更高.

表 2 标准切分和 ICTCLAS 切分对照计算 *SOMLogP*

切分	词序列、1-gram SOM 序列						*SOMLogP*
标准	4	)	**月宫一号**	志愿者	105		**-25.770**
	76	304	**37**	32	76		
ICTCLAS	4	)	**月宫**	**一**	**号**	志愿者	105	-28.904
	76	304	**37**	**79**	**350**	32	76	
标准	实验	后	**出舱**	,	氧气		**-24.208**
	325	116	**346**	390	118		
ICTCLAS	实验	后	**出**	**舱**	,	氧气	-28.209
	325	116	**346**	**230**	390	118	
标准	食物	实现	**自循环**	.	SEND		**-24.630**
	233	105	**117**	395	400		
ICTCLAS	食物	实现	**自**	**循环**	.	SEND	-26.563

3.3 扩展构词规则过滤

基于规则的新词发现研究采用语言学家归纳的一

套构词规则, 如“n+v”、“a+n”等, 在实验中已经证明具有较高的新词识别准确率, 但是规则的总结和维护成本较高[10]. 由于大部分新词依然遵循语言学的构词法[18], 引入构词规则依然有积极作用.

为了提高规则的准确性, 本文在借鉴现有“词性+词性”构词规则基础上提出了“词+词性”的构词规则, 如词“自动”、“自启动”和“自循环”符合 “自+v”的构词规则; 如“实验班”、“实验田”和“实验室”符合“实验+n”的成词规则, 而“天实验”则违反该规则.

为了降低规则总结维护的成本, 本文采用基于词典的构词规则查询过滤方法:

Step1: 对词典 Dict 中的每个词 Word 以字为单位进行全切分, 可获得多个切分序列 Seq_w 组成的集合 Set_w, 然后将集合中非最短切分序列舍弃. 以词“自启动”为例, 可获得切分序列集合{“自/p 启动/v”,“自启/nr2, /wd”,“自/p 启/v 动/v”}, 其中切分序列“自/p 启/v 动/v”被舍弃.

Step2: 计算每个切分序列的概率(取切分单位在词典词语中相应位置出现的概率最小值), 将 Set_w 中概率最大的 Seq_w 作为词 W 的最佳切分 $BestSeq_w$. 词“自启动”的最大概率切分序列为“自/p 启动/v”.

Step3: 获取候选新词的最佳切分序列, 当切分序列长度为 2 时按照 “词+词性”、“词性+词”的形式获得构词组合. 以“天实验”为例其最佳切分为“天/n 实验/v”, 可得到“n+实验”、“天+v”的构词组合.

Step4: 在词典中查找含有候选词切分单位的所有构词模式, 当一种构词模式在词典中出现次数大于 t(t>2)时才被视为规则. 如词典中含有切分单位“实验”的词形成的构词规则只有: “实验+n”.

Step5: 当候选词的构词模式违反构词规则时被过滤.

3.4 新词发现流程

本文提出的新词发现流程如图 3 所示, 主要分为三个阶段:

① 对微博文本进行中文分词. 以分词序列中的单字词为中心, 向左右两侧扩展获取候选字符串, 直到 n 个双字词(n≥1)、1 个停用词[18]或 1 个标点符号为止. 上文例句经过中文分词后, 可获得候选字符串有(n=1): “月宫一号志愿者”、“105 天实验”、“实验后出舱”、“实现自循环”.

② 遍历候选字符串, 对其进行 m(m≥2)字长以内的全切分, 将产生大量候选新词, 但其中包含大量的垃圾词. 以候选串“实现自循环”为例, 将产生 3 字长以内的候选词: “实现”、“现自”、“自循”、“循环”、“实现自”、“现自循”、“自循环”, 其中“实现”为已登录词将被过滤掉.

③ 将每个候选词单独导入分词系统, 对候选字符串进行重新分词获得对应的分词序列, 采用基于 Skip-gram 融合词向量投影的方法计算每个分词序列的 *SOMLogP* 值. 计算候选词分词序列 *SOMLogP* 和原始分词序列 *SOMLogP* 的差值, 从高到低对候选词进行排序, 并结合构词规则对候选词进行过滤.

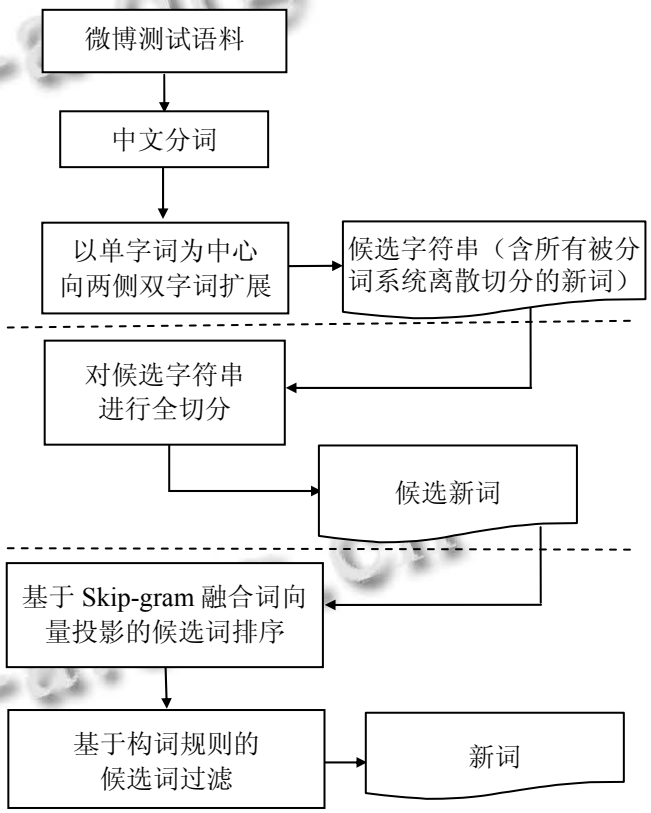

图 3　新词发现流程

如表 3 所示, 每个候选字符串可生成若干候选词, 在排序和规则过滤后每个候选字符串最多可获取一个新词(Δ*SOMLogP* 值最大), 在本例中成功获得全部新词.

表 3　候选字符串中候选词的排序和过滤

候选词	Δ*SOMLogP*(≥1)	违反成词规则
候选字符串: “月宫一号志愿者”		
月宫一号	3.134	
号志愿者	1.788	

一号	1.276	
	候选字符串: “105 天实验”	
~~天实验~~	1.883	“实验”成词规则: 实验+v
	候选字符串: “实验后出舱”	
~~实验后出舱~~	4.808	“实验”成词规则: 实验+v
出舱	4.002	
后出舱	3.666	
	候选字符串: “实现自循环”	
~~实现自循环~~	2.952	“自”成词规则
自循环	1.933	

4 微博新词识别实验

4.1 NLPCC2015 微博语料和评价指标

NLPCC2015[19]的共享任务之一是基于微博语料的中文分词和词性标注. 该任务提供的新浪微博语料包含训练集(已分词的 10000 句微博)、词典(来自训练集)、测试集(未分词的 5000 句微博, OOV Rate=9.75%)和背景数据集(未分词的 58384 行微博语料, 覆盖测试集和训练集). 上述中小规模微博语料也适用于新词发现研究.

新词识别实验采用分词系统 ICTCLAS 2015(训练集词典作为附加词库)对测试集进行预分词处理. 对照 NLPCC 2015 提供的测试集金标准切分, 分词系统本身已经能够正确切分出 57.69%的未登录词(相对于训练集词典), 新词错误切分的情况经统计可划分为以下几种, 如图 4 所示.

a) 新词被拆分, 全部为单字词

b) 新词被拆分, 为单字词和非单字词

c) 新词被拆分, 全部为非单字词

d) 新词和邻近词合并(或部分合并)

其中新词被拆分为含有单字的情况达到了 90.52%, 本文主要对 a、b 这两种新词进行识别.

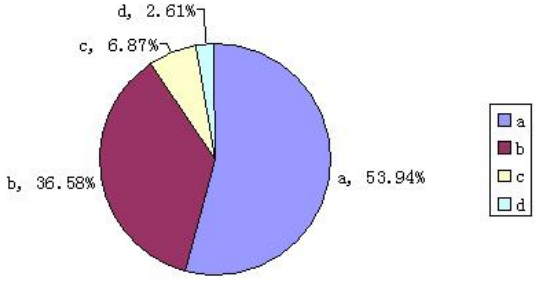

图 4 新词错误切分情况分类

新词识别能力评测标准为新词发现准确率、新词发现召回率、新词发现 F 值, 计算公式如下:

$$precison=\frac{\text{识别正确的新词数目}}{\text{识别的新词数目}} \tag{8}$$

$$recall=\frac{\text{识别正确的新词数目}}{\text{被错误切分的新词数目}} \tag{9}$$

$$F1=\frac{2\times precision\times recall}{precision+recall} \tag{10}$$

4.2 微博语料新词识别测试

使用 word2vec[14]工具对 NLPCC2015 微博训练语料进行训练, 获得 200 维词向量和词性向量模型.

从训练语料生成 1-gram 语料, 并转换为词向量和词性向量, 分别以 0.8 和 0.2 的比重送入 R 平台 kohonen[17]包 20×20 神经元 Super-organizing Maps 进行投影训练, 获得词向量 SOM 映射模型.

从训练语料生成 2-gram 语料, 按上述方法送入 30×30 神经元 Super-organizing Maps 进行训练, 获得映射模型.

使用 ICTCLAS2015 对测试语料进行分词, 依照新词发现流程进行新词识别, 测试结果如表 4:

表 4 新词识别结果

新词发现准确率(%)	新词发现召回率(%)	新词发现 F 值
89.7	68.3	77.6

将识别出的新词 NewWords 作为用户词典导入 ICTCLAS, 对测试语料再次分词. 对照开源分词工具 CRF++、FNLP 和文献[20]GRNN 方法在 NLPCC2015 上的分词实验结果, 如表 5:

表 5 分词结果

分词系统	分词准确率(%)	分词召回率(%)	分词 F 值
ICTCLAS	91.2	92.0	91.7
CRF++	93.3	93.2	93.3
FNLP	94.1	93.9	94.0
GRNN	94.7	94.8	94.8
ICTCLAS+NewWords	94.6	93.8	94.2

5 结语

本文尝试对中小规模微博语料进行低频新词发现, 提出 Skip-gram 融合词向量投影的微博新词发现方法, 能够在实验中有效识别低频新词, 从而提高分词系统的分词成绩.

未来工作可尝试开展对大规模微博语料的新词发现研究.

参考文献

1 黄昌宁,赵海.中文分词十年回顾.中文信息学报,2007,21(3):8−19.

2 Li HQ, Huang CN, Gao JF, Fan XZ. The use of SVM for Chinese new word identification. Natural Language Processing-IJCNLP 2004. Springer Berlin Heidelberg. 2005. 723−732.

3 郑家恒,李文花.基于构词法的网络新词自动识别初探.山西大学学报(自然科学版),2002,25(2):115−119.

4 崔世起,刘群,孟遥,于浩,西野文人.基于大规模语料库的新词检测.计算机研究与发展,2006,43(5):927−932.

5 罗盛芬,孙茂松.基于字串内部结合紧密度的汉语自动抽词实验研究.中文信息学报,2003,17(3):9−14.

6 Feng HD, Chen K, Deng XT, Zheng WM. Accessor variety criteria for Chinese word extraction. Computational Linguistics, 2004, 30(1): 75−93.

7 Li HQ, Huang CN, Gao JF, Fan XZ. The use of SVM for chinese new word identification. Natural Language Processing-IJCNLP 2004. Berlin Heidelberg: Springer-Verlag, 2004: 723−732.

8 Chooi-ling G, Masayuki A, Yuji M. Training multi-classifiers for Chinese unknown word detection. Journal of Chinese Language and Computing, 2005, 15(1): 1−12.

9 Ye YM, Wu QY, Li Y, Chow KP, Hui LCK, Yiu SM. Unknown Chinese word extraction based on variety of overlapping strings. Information Processing & Management, 2013, 49(2): 497−512.

10 霍帅,张敏,刘奕群,马少平.基于微博内容的新词发现方法.模式识别与人工智能,2014,27(2):141−145.

11 廖健,王素格,李德玉,陈鑫.基于构词规则与互信息的微博情感新词发现与判定.第二十届全国信息检索学术会议(CCIR2014).第六届中文倾向性分析评测委员会.昆明.2014.90−96.

12 邱云飞,刘世兴,魏海超,邵良杉.W-POS 语言模型及其选择与匹配算法.计算机应用,2015,35(8):2210−2214.

13 Guthrie D, Allison B, Liu W, Guthrie L, Wilks Y. A closer look at skip-gram modelling. Proc. of the Fifth International Conference on Language Resources and Evaluation. [s.l.]: Conference Publications. 2006. 1222−1225.

14 Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. Proc. of NIPS. [s.l.]: Conference Publications. 2013. 1−9.

15 Bengio Y, Ducharme R, Vincent P, Jauvin C. A neural probabilistic language model. Journal of Machine Learning Research, 2003, 3: 1137−1155.

16 Kohonen T. Self-organizing formation of topologically correct feature maps. Biol Cyber, 1982, 43: 59−69.

17 Ron W, Lutgarde B. Self- and super-organising maps in r: the kohonen. Journal of Statistical Software, 2007, 21(5): 1−19.

18 周超,严馨,余正涛,洪旭东,线岩团.融合词频特性及邻接变化数的微博新词识别.山东大学学报(理学版),2015,50(3):6−10.

19 Qiu XP, Qian P, Yin LS, Wu SY, Huang XJ. Overview of the NLPCC 2015 shared task: Chinese word segmentation and POS tagging for micro-blog texts. Springer International Publishing, 2015, 9362: 541−549.

20 Chen XC, Qiu XP, Zhu CX, Huang XJ. Gated recursive neural network for Chinese word segmentation. Proc. of Annual Meeting of the Association for Computational Linguistics(ACL 2015). The Association for Computational Linguistics. Beijing. 2015.