doi:10.6043/j.issn.0438-0479.201811022

基于子词信息的维吾尔语词项规范化

张新路[1,2],王 磊[1],杨雅婷[1*],米成刚[1]

(1.中国科学院新疆理化技术研究所,新疆民族语音语言信息处理实验室,新疆 乌鲁木齐 830011;
2.中国科学院大学计算机科学与技术学院,北京 100049)

摘要:拉丁化的维吾尔语在使用过程中具有文本不规范的特点,这种不规范是造成歧义等现象的最主要原因,严重制约着与维吾尔语相关的自然语言处理应用.由此提出了一种无监督的基于子词信息的文本规范化方法,该方法在词向量构建过程中将词的内部信息考虑进去.这种方法可以对罕见词进行向量表示,也可以将词内部的形态信息融入词的表示,丰富词向量的表达,进而用于改进无监督学习中规范化词候选集生成质量的不足.实验表明,相比于传统词向量构建方法,该方法在文本规范化任务中可以提高规范化词的召回率.

关键词:维吾尔语;自然语言处理;文本规范化;词嵌入

中图分类号:TP 391 **文献标志码**:A **文章编号**:0438-0479(2019)02-0217-08

随着社交媒体的普及,各社交平台每天都会在网络上产生海量的文本数据,例如当下的时事热点、产品的评价、舆论的监控、实时灾情报告等.这些文本数据所蕴含的信息对社会的诸多领域都有着极高的应用价值.要从这些文本中提取相应的信息,首先要进行分词、词性标注[1]、句法分析[2]等一系列自然语言处理操作,但是文本的不规范性所造成的数据稀疏严重影响了自然语言处理操作的有效性[3].在汉语、英语等广泛使用的社交媒体语言中,通过规则替换、噪声信道模型以及基于神经网络编码、解码等模型对不规范文本处理有较为深入的研究;但在维吾尔语中,由于语料资源的匮乏以及维吾尔语词形变化的多样性,不规范文本处理的准确性还存在较大的差距.因此,研究如何更好地分析与处理维吾尔语的不规范文本信息是非常有意义和价值的[4].

维吾尔语是在中国境内广泛使用的社交媒体小语种,在形态结构上属于黏着性语言[5].维吾尔语的书写中采用阿拉伯字母、拉丁字母以及西里尔字母等形式,目前广泛使用的是基于阿拉伯字母的书写方式,然而在某些情况下,用户为了克服阿拉伯字母限制会使用基于拉丁字母的书写方式[6],简称拉丁维语.由于拉丁维语在早期发展过程中没有统一的标准,在基于阿拉伯字母的老维语与拉丁维语转换的过程中,有时会因为书写习惯的不同而产生不规则的情况.

一项关于维吾尔族民众使用拉丁字母情况的调查显示,170 名受访者中有 39.8%不使用拉丁字母,29.7%使用标准的拉丁字母,还有 30.5%使用不规则的拉丁字母,例如“X”、“SH” 和“Ş”都可能作为老维语字母ش的转换[7].在老维语的 32 个字母中有 15 个存在 2~4 个可能的拉丁字母转换[7],这 15 个字母的转换方式如表 1 所示.这种不规则转换会产生严重的文本歧义现象(如表 2 所示),给后续自然语言处理操作的有效性带来很大的影响.因此,本研究主要针对拉丁维语中的不规则转换来进行文本规范化处理.

文本规范化任务主要分为有监督的方法、无监督的方法以及基于规则转换的方法.维吾尔语属于形态丰富的语言,词的词汇和语法变化均通过在实词词干上缀接各种附加成分实现[8],这种特性使得维吾尔语具有词汇量过大的特点,基于规则的方法很难在维吾

收稿日期:2018-11-12 **录用日期**:2018-12-05

基金项目:国家自然科学基金(U1703133);新疆自治区重大科技专项(2016A03007-3);中国科学院“西部之光”人才培养引进计划(2017-XBQNXZ-A-005)

* **通信作者**:yangyt@ms.xjb.ac.cn

引文格式:张新路,王磊,杨雅婷,等.基于子词信息的维吾尔语词项规范化[J].厦门大学学报(自然科学版),2019,58(2):217-224.

Citation:ZHANG X L,WANG L,YANG Y T,et al. Normalization of Uyghur terms based on subword information[J]. J Xiamen Univ Nat Sci,2019,58(2):217-224.(in Chinese)

表 1 老维语字母转换为拉丁字母的可替换形式

Tab. 1 Possible Latin alphabet alternatives of Uyghur Perso-Arabic alphabet

维吾尔字母	ژ	ئو	ئە	ق	ئى	ئۇ	چ	ئۆ	غ	خ	ش	ھ	ڭ	ئې	ۋ
标准形式	J	O	E	Q	I	Ü	Ç	Ö	Ğ	X	Ş	H	Ñ	É	V
可替换形式	ZH, J, Z, Y	O U	A E	Q K	I E	V, U, O, Ü	Ç, Q, CH	O,U, V, Ö	G,GH, H,Ğ	X H	X, SH, Ş	H Y	G, NG, Ñ	E, I, Ë, É	V W

表 2 不规则替换所产生的歧义现象

Tab. 2 Ambiguity caused by irregular substitution

标准形式	不规则替换所产生的歧义
Qan(血液)	Kan(矿井)
Söz(单词)	Soz(伸展)
Oruq(瘦的)	Örük(杏黄色)
Söyuş(亲吻)	Soyuş(剥离)
Kelgin(过来)	Qalghin(停留)

尔语上取得较好的效果. 有监督的学习方法需要在大规模的标注语料上进行训练,对于维吾尔语这种小语种而言,很难获取到大规模的标注语料,因此语料规模限制了有监督的方法在维吾尔语上的应用. 越来越多的研究者通过无监督的方法来进行文本规范化. 无监督的方法要对词生成向量表示,然而常用的向量表示方法只是对每个词生成独立的向量表示,忽略了词的内部形态[9],因此这种表示方式对于词汇量巨大并且有很多罕见词的维吾尔语来说有很大的限制.

本研究针对维吾尔语在无监督学习中所产生的规范化词候选集质量的不足,引入了子词信息,并在 Sridhar[10] 的基础上进一步优化候选词生成的质量,用于维吾尔语的文本规范化,提出了一种基于子词信息的维吾尔语文本规范化的方法. 该方法的改进主要体现在以下两个方面:

1) 在文本规范化任务中,许多单词由于字母的错写、漏写、缩写以及词干和词缀的组合多样性会产生大量的罕见词,但是传统的词嵌入模型通过统计词的共现频率来得到相应的向量表示,许多罕见词由于共现频率小而得不到相应的向量表示. 本文中提出的方法对罕见词进行字符串切分,通过字符串的向量组合获得罕见词的向量表示.

2) 一般的词向量表示只是对每个词产生一个独立的向量,忽略了词内部的形态结构. 本文中提出的方法通过表征隐藏在类别间的共享信息,对词之间的形态相似性有更好的表达. 对于文本规范化任务来说,考虑词之间形态相似性的向量表示会有更好的效果.

1 相关工作

文本规范化是语料预处理的一个关键步骤,一直以来都吸引了很多研究者的目光. 随着噪声文本的指数增长,文本规范化已成为自然语言处理领域的研究热点.

Brill 等[11] 在 2000 年将噪声信道模型应用于文本规范化中,之后 Toutanova 等[12] 对噪声信道模型进行扩展,将词的发音作为特征加入模型中. 这种模型在特定范围内能够取得较好的文本规范化效果,但属于有监督的模型,需要大量的标注语料对模型进行训练.

由于之前的文本规范化方法都是基于词的一对一映射,对于文本规范化任务来说有很大的限制,因此产生了基于统计机器翻译的方法. 文本规范化可以被看作是将不规范文本翻译成规范化文本的单语言机器翻译过程. 借助于机器翻译方法中的词对齐概念,可以对非标准词与标准词关系中的一对多、多对一和多对多映射进行建模. Aw 等[13] 在 2006 年提出了基于词组的机器翻译方法;由于更细粒度的对齐方式可以更好地捕捉到单词变形的共性,Pennell 等[14] 提出了基于字符的机器翻译方法;随着神经网络技术的不断发展,Xie 等[15] 将基于注意力机制的字符级序列模型用于语言校正;Ikeda 等[16] 使用了一种神经网络编码、解码模型,在日语文本中通过 3 种不同的书写系统来实现噪声的正常化. 然而与噪声信道模型类似,上述方式在训练阶段需要大量的训练数据,对于维吾尔语这种语料匮乏的小语种来说,很难大规模获取这种一一对应的训练语料.

由于语料规模的限制,很多研究者将目光转向无监督学习的方法上. Han 等[17] 提出利用训练语料的上下文信息产生非标准词与标准词相对应的词对,进而产生文本规范化词典,用于微博文本规范化;Hassan 等[18] 提出基于二分图的无监督随机游走算法,用于社

交媒体上的文本规范化,该方法把两个词的上下文相关性当作两个词的相关性依据,从而用于归一化. Sridhar[10]将词的分布表示用于文本规范化,该方法利用特定词的噪声和规范形式在一个大的噪声文本集合中共享类似的上下文这一特性,通过单词的分布式表示捕获上下文相似的概念,并以完全不受监督的方式从这些表示中学习规范化词汇表,但是该方法在维吾尔语这种形态丰富的语言中,由于不能产生较好的规范化词候选集,所以对于文本规范化的准确性有较大的影响.

本文在 Sridhar[10]的基础上进一步优化词表示的方法,将词的形态信息及字符级信息用于维吾尔语词的向量表示,以提高维吾尔语文本规范化的准确性.

2　基于子词信息的表示方法

本研究主要针对拉丁维语中的不规则转换来进行文本规范化处理. 基于维吾尔语本身的特点,文本规范化模型主要分为以下两个步骤:首先对于含有不规则转换的语料生成词嵌入表示,在这一部分提出了基于子词信息的词表示方法,该方法通过计算子词单位信息来构建词法模型,使用字符级的 n-元语法(n-gram)之和表示词,将词信息映射到向量空间模型;然后对于含有不规则转换的词,通过词之间的余弦相似度选择与之最相近的词作为候选词.

2.1　基于子词信息的分布表示

词是承载语义的最基本单元,而最初的词独热(one-hot)表示仅仅将单词符号化,不包含任何语义信息. Harris[19]提出"上下文相似的词,其语义也相似"的分布假说,为将语义融入词表示中提供了理论基础.

Mikolov 等[20]提出的 word2vec 工具能够基于海量文本库和无监督策略实现高效的词向量学习. 如图 1 所示,连续词袋(CBOW)和跳字(Skip-gram)模型是 word2vec 工具中用于描述词间共现关系的统计模型,其中 V 表示词表(词的集合),w_i 表示文本序列中第 i 个词,m 表示上下文词的范围. 设计 2 个模型主要是希望用更高效的方法获取词向量.

在 word2vec 工具中针对每个词会生成不同的向量表示,然而这种表示会忽视词的内部形态结构,因此本研究在这些模型的基础上将词的 n-gram 考虑进去. 以拉丁化的维吾尔语单词 médiyeliri(赞赏)为例,在实验中选择 n 的范围为 3～6,将每个词切分成大小

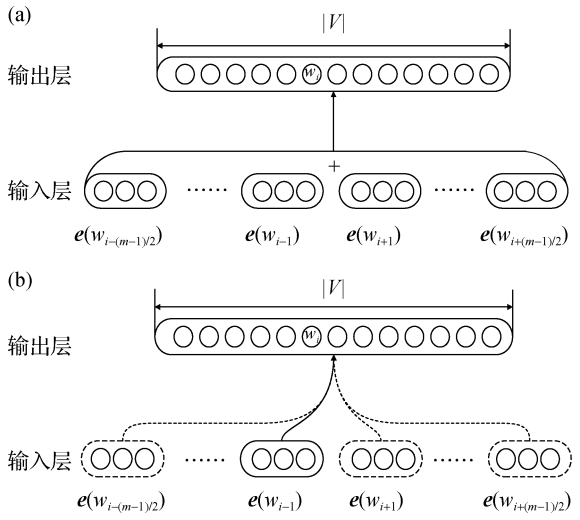

图 1　CBOW(a)和 Skip-gram(b)模型结构图[20]

Fig. 1　Structure diagrams[20] of CBOW (a) and Skip-gram (b) models

长度为 n 的子字符串,则它的 n-gram 有如下几种形式:

$n=3$　"＜mé", "méd", "édi", "diy", "iye", "yel", "eli", "lir", "iri", "ri＞";

$n=4$　"＜méd", "médi", "édiy", "diye", "iyel", "yeli", "elir", "liri", "iri＞";

$n=5$　"＜médi", "médiy", "édiye", "diyel", "iyeli", "yelir", "eliri", "liri＞";

$n=6$　"＜médiy", "médiye", "édiyel", "diyeli", "iyelir", "yeliri", "eliri＞".

同时还要将词本身考虑在内,即

＜médiyeliri＞,

其中,＜表示前缀,＞表示后缀. 可以用这些范围内的 n-gram 向量叠加来表示 médiyeliri 这个词.

对于给定的语料库,设基于词表所获得的词 w 的字符级n-gram的规模大小为 G,则给定一个词 w,可以通过 $g_w \subset \{g_1, g_2, \cdots, g_G\}$ 将词 w 表示为 n-gram 的字母组合,将每个 n-gram 的字符 g 生成一个向量表示 $\boldsymbol{z}_g$;其上下文 c 的向量表示为 $\boldsymbol{v}_c$,可以通过对词 n-gram 的向量与上下文向量的乘积进行求和来表示一个词. 因此得到一个得分函数:

$$s(w,c) = \sum_{g \in g_w} \boldsymbol{z}_g^{\mathrm{T}} \boldsymbol{v}_c. \tag{1}$$

将得分函数应用于 CBOW 和 Skip-gram 模型词向量的生成(图 2). 为了限定模型的内存要求,使用散列函数将字符级的 n-gram 映射到 1～K 的整数中,采用 Fowler-Noll-Vo 散列函数来映射字符序列,并将 K 设置为 2×10^6. 每个词根据它在单词字典中的索引及其包含的哈希值的集合来获得相应的表示.

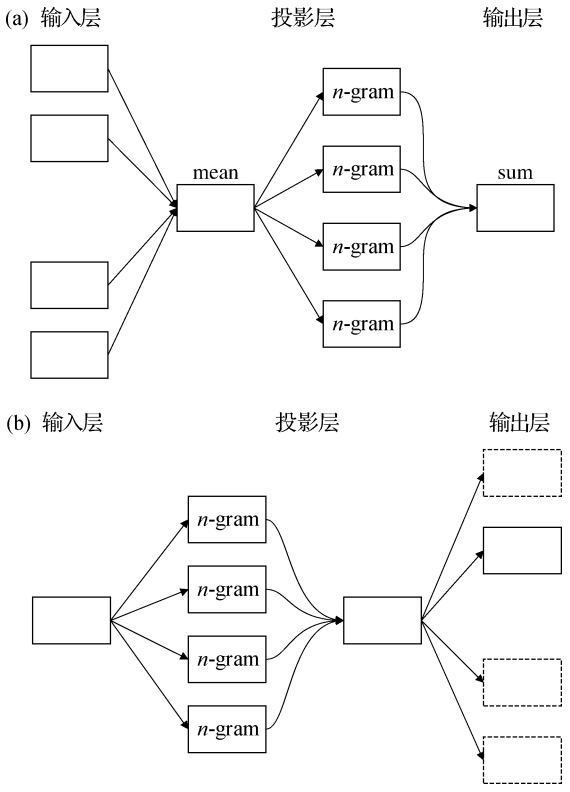

图 2 基于子词信息的 CBOW (a)和 Skip-gram (b)模型

Fig. 2 CBOW (a) and Skip-gram (b) models based on subword information

CBOW 模型如图 2(a)所示,使用一段文本的中间词作为目标词,在神经网络语言模型的基础上做了两方面的简化:1) 不包含隐藏层,该模型从神经网络结构直接转化为对数线性结构,与 logistic 回归一致,对数线性结构比三层神经网络结构少了一个矩阵运算,大幅度地提升了模型的训练速度;2) 去除了上下文各词的词序信息,使用上下文各词词向量的平均值代替神经网络语言模型使用的上下文各词词向量的拼接[21]. 对于一段训练文本 $w_{i-(m-1)/2},\cdots,w_i,\cdots,w_{i+(m-1)/2}$,则输入层为

$$\boldsymbol{x}=\frac{1}{m-1}\sum_{w_j\in c}\boldsymbol{e}(w_j), \tag{2}$$

其中 $\boldsymbol{x}$ 为上下文词向量的表示,即为前文中的 $\boldsymbol{v}_c$.

设目标词 w 和上下文 c 分别为

$$w=w_i, \tag{3}$$

$$c=\{w_{i-(m-1)/2},\cdots,w_{i-1},w_{i+1},\cdots,w_{i+(m-1)/2}\}, \tag{4}$$

CBOW 模型的输入层为上下文的表示,则可根据上下文的表示直接对目标词进行预测:

$$P(w\mid c)=\frac{\exp(s(w,\boldsymbol{x}))}{\sum_{w'\in V}\exp(s(w',\boldsymbol{x}))}, \tag{5}$$

其中 w'为词表中的词.

对于整个语料而言,该模型的优化目标为最大化

$$\sum_{(w,c\in D)}\log P(w\mid c), \tag{6}$$

其中 D 表示训练词向量的语料.

Skip-gram 模型如图 2(b)所示,与 CBOW 模型一致,该模型中也没有包含隐藏层;而与 CBOW 模型不同的是,该模型每次从 w 的 c 中选择一个词,将其词向量作为模型的输入 x,即上下文的表示. 由于模型需要遍历整个语料,任意一个窗口中的两个词 w_a 和w_b 都需要计算 $P(w_a\mid w_b)+P(w_b\mid w_a)$,因此 Skip-gram 模型同样是通过上下文预测目标词[22],对于整个语料的优化目标为最大化

$$\sum_{(w,c)\in D}\sum_{w_j\in c}\log P(w\mid w_j), \tag{7}$$

其中,

$$P(w\mid w_j)=\frac{\exp(s(w,w_j))}{\sum_{w'\in V}\exp(s(w',w_j))}. \tag{8}$$

在词向量的表示中,融入子词信息可以学习到词内部的形态知识,当两个词有相同的词缀时,其语义也具有一定的相似性. 例如英语单词 disagree 和 discover,从词层面上看是两个不同的词,但是如果用字符级的 n-gram 来表示,则有相同的前缀 dis 来表示否定的含义. 对于维吾尔语这种形态丰富的语言来说,这种方式可以更好地表现出词的含义.

2.2 相似性的度量

通过上述模型将词映射到向量空间,形成文本中文字和向量数据的映射关系,获取词向量和概率密度函数. 词向量是多维实数向量,向量中包含了自然语言中的语义和语法关系,对于文本规范化任务来说向量的绝对长度并不重要,重要的是两个向量之间的角度. 余弦距离更多是从方向上区分差异,而对绝对数值不敏感,词向量之间余弦距离的大小代表了词之间关系的远近. 因此,本文中通过余弦相似度来衡量两个词之间的相似性,定义如下:

$$\mathrm{sim}(u_i,v_i)=\frac{\sum_{i=1}^{M}u_i\times v_i}{\sqrt{\sum_{t=1}^{M}(u_i)^2\times\sum_{i=1}^{M}(v_i)^2}}, \tag{9}$$

其中,u_i,v_i 分别是两个词向量的第 i 维分量,M 表示词向量的维度.

3 实验与分析

3.1 数据集

3.1.1 训练数据集

从 www.tianshannet.com,www.okyan.com 以

及 www.uycnr.cn 等维吾尔语网站上爬取文本信息作为训练数据,共收集了约 2 GB 的文本文件.这些文本文件都是基于阿拉伯字母的老维语,由于老维语和拉丁维语的转换具有相应的规范,通过这种规范将老维语转换为拉丁维语,获取到规范的拉丁维语训练数据.

一项关于在使用维吾尔语写作时如何使用拉丁字母的调查中,发现了一些关于拉丁字母不规则替换的规律[7],如表 3 所示.

表 3　拉丁字母可能的替换形式

Tab.3　Possible alternatives of Latin characters

字符	可替换	字符	可替换	字符	可替换
u	u,ö,ü	a	a,e	ë	é
v	v,ö,ü	k	k,q	ch	ç
o	o,ö,ü	n	n,ñ	ng	ñ
i	i,é	j	j,c	sh	ş
h	h,x,ğ	q	q,ç	gh	ğ
y	y,h,j	x	x,ş	zh	j
e	e,é,i	k	k,q	ë	é
g	g,ñ,ğ	z	z,j		

由于大规模不规范文本的训练数据很难获取,因此参考表 3 中的规则,使用随机函数 rand 将当前时间作为随机数种子,随机选一个可替换字母替换源文本中的字母,人工生成拉丁字母不规则使用的训练文本.由于在替换表中存在替换字符长度不相等的情况,例如 zh 可能替换为 j,对此在被替换的词后面添加字母 w(字母 w 在标准字母表中不存在,不影响语义的表达),以确保源文本和目标文本的长度一致.

3.1.2　测试数据集

从维吾尔语新闻网站天山网上爬取一些文本信息作为测试数据.为了更好地评价基于词的子信息在不同类别文本上的效果,采用其中的 5 个短文本来进行文本规范化测试,每个文本信息平均包括 566 个词,具体信息如表 4 所示,其中 OOV(out of vocabulary)词表示集外词,这些文本涵盖了新闻、文学、体育、文化等领域.采用训练数据的处理方式,可以得到与规范化文本相对应的噪声文本.

3.2　参数设置

实验中有多个参数需要人工设置,其中 word2vec 工具中词向量的维数经实验发现与语料库的大小有关,在实验中将其设置为 100,上下文窗口大小为 5,最低词频为 5;采用负采样技术来提升最后一层的效率,采样的阈值大小为 5;对于产生的规范化文本候选集,设定其大小为 10.基于词的子信息的词向量参数设置与 word2vec 工具中一致.对于每个词的 n-gram 切分,选择 n 为 3～6,即对每个词切分成长度为 3～6 的子字符串,每个词表示为这些子字符串的向量之和.该方法和基线系统共享相同的词表,因此在相同训练集中不同方法的结果是可以比较的.

表 4　测试文本信息

Tab.4　Test text information

文本编号	总词数	OOV	
		词数	百分比/%
1	560	91	16.25
2	521	34	6.53
3	712	20	2.81
4	494	73	14.78
5	544	40	7.35

3.3　评价指标

对于每个不规范词,根据余弦相似度产生 10 个最相近的规范化候选词.为了提高规范化候选词生成质量,以召回率,即正确规范词数占所需规范总词数的百分比,作为衡量文本规范化准确性的指标.同时采用斯皮尔曼等级相关系数(Spearman's rank correlation coefficient,ρ)作为词汇相似度的评价指标,它是衡量两个变量依赖性的指标,利用单调方程评价两个统计变量的相关性[21].如果数据中没有重复值且当两个变量完全单调相关时,其值则为 +1 或 −1;对于样本容量为 n 的样本,ρ 的计算公式如下:

$$\rho=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2\sum_{i=1}^{n}(Y_i-\overline{Y})^2}},\quad(10)$$

其中,X_i 表示两个词之间的余弦相似度分数,Y_i 表示人工标注的相似度分数.

3.4　结果与分析

3.4.1　规范化词召回率

用 word2vec 工具在训练语料中生成相应的词向量表达,但是由于 word2vec 工具并不能对 OOV 词进行向量表示,所以排除这些词后进行评测.首先对测试文本中的 OOV 词进行了统计,通过表 4 中的结果可以发现,在 5 个测试文本中平均约有 9% 的 OOV 词.

由于词向量生成过程中具有两种不同的上下文

表示,所以对这两种表示方式分别做了实验对比. Skip-gram 模型中两种词向量生成方式对于规范化词的召回率如图 3 所示. 相对于 word2vec 工具,本研究提出的方法获得的平均召回率提高了 3.5 个百分点,但是在第 4 个测试文本中两种方法产生相同的召回率,这说明 Skip-gram 模型并不能保证在所有文本上都取得较好的结果. 对于产生这种现象的原因从模型结构上进行了分析:该模型将窗口中的一个词作为上下文的表示,在小规模的语料上会取得较好的结果;对于更大规模的语料,这种建模方式不能很好地体现出上下文之间的关系,在引入词的子信息的词向量表示时,则会限制词向量的表示. 因此,该模型在文本规范化任务上并不能得到较好的提升.

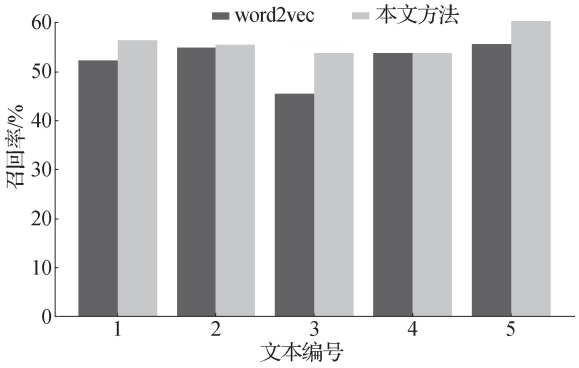

图 3 基于 Skip-gram 模型的召回率

Fig. 3 Recall rate based on Skip-gram model

进而在更复杂的上下文表示方式的 CBOW 模型上进行了实验,结果如图 4 所示. 相对于 word2vec 工具,本研究提出的方法在所有测试文本上都取得了较好的结果,平均召回率提高了 8 个百分点. 由于 CBOW 模型采用了更复杂的上下文表示,引入词的子信息能够进一步丰富词向量的表示,所以在大规模语料上所生成的词向量能更好地体现出上下文之间的关系. 对于文本规范化任务来说,通过词向量空间距离最近的词来生成规范化词的候选集,这种方法能够取得更高的召回率.

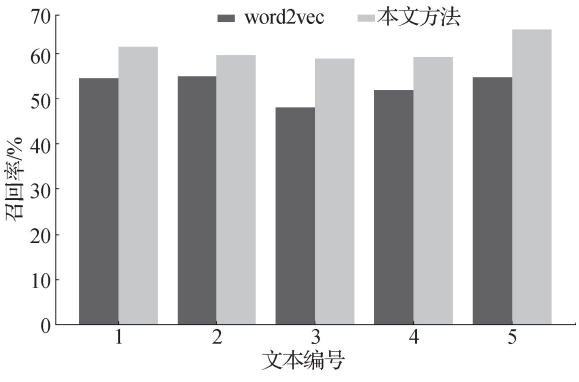

图 4 基于 CBOW 模型的召回率

Fig. 4 Recall rate based on CBOW model

由于引入词的子信息可以获得罕见词的表示,所以对于 word2vec 工具中的 OOV 词也可以获得其相应的向量表示,进而构建规范化词的候选集. 对 OOV 词的召回率进行了统计,结果如表 5 所示.

表 5 对于 OOV 词的召回率

Tab. 5 Recall rates for OOV words %

文本编号	召回率	
	Skip-gram	CBOW
1	1.1	1.1
2	0	0
3	10.0	10.0
4	1.4	1.4
5	2.5	0

通过对 OOV 词的分析,发现引入词的子信息虽然可以对 OOV 词进行表示,但是并不能产生很好的效果,对这一现象进行了如下分析:

1) 在测试文本中的 OOV 词是一些缩写或者合成词,通过词的子信息可以找到与之形态相似的词,例如标准文本为"orda-saraylarğa"(房屋群),噪声文本为"orda-saraylargha",基于词的子信息找到与之相似的词为"soda-saraylarğa",噪声文本字符串被正确恢复,但是由于组合的多样性,未正确恢复到规范化的形式,所以对于这种文本的恢复还需要改进.

2) 由于计算方式的差别,这些 OOV 词没有出现在单词字典中,在计算词向量的方式上没有这些词的完整词向量表达,所以计算这些词时不能完整地表达出该词的含义,只能根据字符的 n-gram 来计算向量表达,由此导致在词恢复的过程中不能更好地利用上下文相似的属性.

3.4.2 ρ 值

为了能够更准确地评价词汇之间的相似度,人工标注了 50 组词汇的相似度系数,范围为 0～5,相关性越高,分值越大. 通过人工标注的相似度与词向量所产生的余弦相似度计算 ρ 值,结果表明在 word2vec 工具上 $\rho=0.48$,而使用本文中的方法在该数据集上 $\rho=0.50$,提高了 0.02,说明引入子词信息在维吾尔语中能够更好地体现词之间的相似性. 这可能是因为一部分维吾尔语单词是复合型词汇,很多单词共享相同的

词根、词缀,词形相似时语义也有一定程度的相关性,所以引入子词信息能够更好地体现出这种相关性.

4 结　论

本研究针对维吾尔语文本规范化任务中候选词质量的不足,引入了词的子信息来丰富维吾尔语的词向量表达.该方法在词向量构建过程中将词进行字符串切分,获取词内部的字符串信息,可以进一步表示词之间的相似性.实验表明,在上下文表示更丰富的模型中引入子词信息可以在规范化词的召回率上提升 8 个百分点.该方法对于 OOV 词的处理还存在一定不足,接下来需要进一步完善,采用更合理的方式来构建相应的模型,以提高 OOV 词恢复的效果.

参考文献:

[1] GIMPEL K, SCHNEIDER N, O'CONNOR B, et al. Part-of-speech tagging for Twitter: annotation, features, and experiments [C] // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland: Association for Computational Linguistics, 2011: 42-47.

[2] FOSTER J, ÖZLEM, WAGNER J, et al. From news to comment: resources and benchmarks for parsing the language of web 2.0[C]// International Joint Conference on Natural Language Processing. Chiang Mai: Asian Federation of Natural Language Processing, 2011: 893-901.

[3] RITTER A, CHERRY C, DOLAN B. Unsupervised modeling of Twitter conversations [C] // Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles: Association for Computational Linguistics, 2010: 172-180.

[4] 罗程多,吴晓蕊,薛凯,等.社交文本规范化研究综述[J].网络新媒体技术,2017(5):10-14.

[5] MI C, YANG Y, ZHOU X, et al. A phrase table filtering model based on binary classification for Uyghur-Chinese machine translation [J]. Journal of Computers, 2014, 9(12): 2780-2786.

[6] 杨帆.新疆维吾尔族大学生手机短信交际语言文字使用现状调查研究[D].乌鲁木齐:新疆师范大学,2012:28-31.

[7] TURSUN O, CAKICI R. Noisy Uyghur text normalization [C]// Proceedings of the 3rd Workshop on Noisy User-generated Text. Copenhagen: Association for Computational Linguistics, 2017: 85-93.

[8] 罗延根,李晓,蒋同海,等.基于词向量的维吾尔语词项归一化方法[J].计算机工程,2018(2):220-225.

[9] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information [EB/OL]. [2018-11-11]. https://arxiv.org/pdf/1607.04606.

[10] SRIDHAR V K R. Unsupervised text normalization using distributed representations of words and phrases[C]// Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. Denver: Association for Computational Linguistics, 2015: 8-16.

[11] BRILL E, MOORE R C. An improved error model for noisy channel spelling correction[C]// Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Hong Kong: Association for Computational Linguistics, 2000: 286-293.

[12] TOUTANOVA K, MOORE R C. Pronunciation modeling for improved spelling correction[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002: 144-151.

[13] AW A T, ZHANG M, XIAO J, et al. A phrase-based statistical model for SMS text normalization [C] // Proceedings of the COLING/ACL on Main Conference Poster Sessions. Sydney: Association for Computational Linguistics, 2006: 33-40.

[14] PENNELL D, LIU Y. A character-level machine translation approach for normalization of sms abbreviations[C]// Proceedings of 5th International Joint Conference on Natural Language Processing. Chiang Mai: Asian Federation of Natural Language Processing, 2011: 974-982.

[15] XIE Z, AVATI A, ARIVAZHAGAN N, et al. Neural language correction with character-based attention[EB/OL]. [2018-11-11]. https://arxiv.org/pdf/1603.09727.

[16] IKEDA T, SHINDO H, MATSUMOTO Y. Japanese text normalization with encoder-decoder model [C] // Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT). Osaka: Association for Computational Linguistics, 2016: 129-137.

[17] HAN B, COOK P, BALDWIN T. Automatically constructing a normalisation dictionary for microblogs[C]// Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island: Association for Computational Linguistics, 2012: 421-432.

[18] HASSAN H, MENEZES A. Social text normalization using contextual graph random walks[C]// Proceedings

of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia: Association for Computational Linguistics,2013:1577-1586.

[19] HARRIS Z S. Distributional structure[J]. Word,1954,10(2/3):146-162.

[20] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[EB/OL]. [2018-11-11]. https://arxiv.org/pdf/1301.3781.

[21] 陈培,景丽萍. 融合语义信息的矩阵分解词向量学习模型[J]. 智能系统学报,2017(5):83-89.

[22] 来斯惟. 基于神经网络的词和文档语义向量表示方法研究[D]. 北京:中国科学院大学,2016:5-25.

Normalization of Uyghur terms based on subword information

ZHANG Xinlu[1,2],WANG Lei[1],YANG Yating[1*],MI Chenggang[1]

(1. Xinjiang Laboratory of Minority Speech and Language Information Processing,the Xinjiang Technical Institute of Physics & Chemistry,Chinese Academy of Science,Urumqi 830011,China;
2. School of Computer Science and Technology,University of the Chinese Academy of Sciences,Beijing 100049,China)

Abstract: Latinized Uyghur language is characterized by nonstandard text in its use. This kind of non-standard type primarily causes the ambiguity,which seriously restricts the application of natural language processing related to Uyghur. This paper proposes a text normalization method based on subword information. The method takes the internal information of words into account in the process of constructing word vectors. In this way,rare words can be represented by the vector,and the morphological information inside the words can also be incorporated into the expression of the words to enrich the expression of the word vectors,which can be used to improve the quality of standardized word candidate set generation. Experimental results show that the proposed method can improve the recall rate of normalized words in text normalization tasks compared with traditional word vector construction methods.

Keywords: Uyghur;natural language processing;text normalization;word embedding