·人工智能及识别技术·　　文章编号：1000-3428(2018)02-0220-06　　文献标志码：A　　中图分类号：TP391

基于词向量的维吾尔语词项归一化方法

罗延根[1,3]，李　晓[1,2]，蒋同海[1,2]，杨雅婷[1,2]，周　喜[1,2]，王　磊[1,2]

（1. 中国科学院新疆理化技术研究所，乌鲁木齐 830011；
2. 中国科学院新疆民族语音语言信息处理重点实验室，乌鲁木齐 830011；3. 中国科学院大学，北京 100049）

摘　要：使用无监督的方法，将口语文本中的非正规维吾尔语词项归一化到正规文本中意思相近的正规词，基于神经网络，利用大规模语料将维吾尔语单词映射到低维向量空间，对向量空间的非正规词进行聚类。引入一个贪心解码器对非正规词做归一化处理，并进行重采样迭代，从而将之前未能成功归一化的非正规词归一化。实验结果表明，使用该方法对维汉机器翻译的待翻译口语文本进行前编辑后，生成的译文质量有显著提高。该方法给维汉口语文本机器翻译系统提供一个前处理的流程，在缺乏双语口语平行语料的情况下也能有效提高机器翻译系统性能。

关键词：维吾尔语口语文本；非正规词；归一化；神经网络；重采样

中文引用格式：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，2018，44(2)：220-225.

英文引用格式：LUO Yan'gen，LI Xiao，JIANG Tonghai，et al. Uyghur Lexicon Normalization Method Based on Word Vector[J]. Computer Engineering，2018，44(2)：220-225.

Uyghur Lexicon Normalization Method Based on Word Vector

LUO Yan'gen[1,3]，LI Xiao[1,2]，JIANG Tonghai[1,2]，YANG Yating[1,2]，ZHOU Xi[1,2]，WANG Lei[1,2]

（1. The Xinjiang Technical Institute of Physics and Chemistry，Chinese Academy of Science，Urumqi 830011，China；
2. Xinjiang Laboratory of Minority Speech and Language Information Processing，
Chinese Academy of Science，Urumqi 830011，China；3. University of Chinese Academy of Sciences，Beijing 100049，China）

【Abstract】A unsupervised approach to normalize the irregular Uygur words in the spoken text to normal words in the formal text. Based on neural network，Uygur words are mapped to a low dimensional vector space by using a large corpus. The irregular words in vector space are clustered. A greedy decoder is introduced to normalize the unformal words and to resample iterations，so as to normalize the unformal words which have not been successfully normalized before. Experiment results show that using this approach to pre-edit the text to be translated by Uyghur-Chinese machine translation，the quality of the generated translation is significantly improved. This method provides a pretreatment process to spoken text and machine translation system，which can effectively improve the system performance of machine translation in the absence of bilingual parallel corpus of spoken.

【Key words】Uyghur spoken text；unformal word；normalization；neural network；resample

DOI：10.3969/j.issn.1000-3428.2018.02.038

0　概述

词汇归一化是将看起来不完全一致的多个词条归纳成一个等价类，是众多自然语言处理方面前处理的一个重要步骤。大部分自然语言处理的工作都要求在一个限定的词汇表上进行处理，这样能够降低模型的复杂度。例如机器翻译、命名实体抽取、信息检索等研究，它们处理的数据都是经过归一化之后的“干净”语料。

近年来，随着互联网的发展，社交媒体上的文本也呈爆炸式增长，但是社交媒体上用户产生的文本并不是很正规的文本，它包含很多不合语法的句子、不正规拼写的单词等。对于这种文本进行自然语言处理的相关工作就显得特别困难，因为有太多的未

基金项目：新疆维吾尔自治区青年科技创新人才培养工程项目(2014711006，2014721032)；新疆维吾尔自治区高技术研究与发展项目(201412101)；新疆维吾尔自治区重点实验室开放课题“基于黏着语形态特征的维汉机器翻译最大熵调序研究”(2015KL031)；新疆维吾尔自治区重大科技专项课题“维汉机器翻译平台”(2016A03007-2)。

作者简介：罗延根(1992—)，男，硕士研究生，主研方向为机器翻译、自然语言处理；李　晓、蒋同海，研究员、博士；杨雅婷，副研究员、博士；周　喜、王　磊，研究员、博士。

收稿日期：2016-12-23　　**修回日期：**2017-02-23　　**E-mail：**yangyt@ms.xjb.ac.cn

登录词。在对用户产生的文本进行处理之前,词汇归一化就显得特别重要。

本文提出将含有非正式维吾尔语用语的社交媒体语料与正规维吾尔语用语的新闻语料结合起来用于获取一个词的低维向量空间,将正规用语的语料中的词汇当作候选词,对于向量空间中的集外词(OOV),首先找到向量空间中的 k 近邻,再对 k 近邻的正规词进行相似度筛选,最后选出一个 *n*-best 的候选词汇集。对于非正规语料句子中的集外词(非正规词),从候选词汇集中选出一个最优的对应的词,类似于机器翻译的解码过程,采用贪心解码器,评估指标为综合字符串相似度以及语言模型的一个评分。

1　相关工作

维吾尔语在形态结构上属于粘着语类型,作为粘着语类型的语言,词的词汇变化和各种语法变化都是通过在实词词干上缀接各种附加成分的方式来表现的,习惯于词干加上后缀去表达不同的含义,例如人称、数量、词态及语气等。维吾尔语由阿拉伯字母组成,字母的错写、漏写、缩写以及词干词缀组合的多样性也导致了维吾尔语中词汇量过大的现象,从而造成严重的数据稀疏性。在大词汇量的基础上衍生出来的非正规词的数量更是庞大,因此,日常用语(非新闻等官方用语)的机器翻译所面临的集外词(Out of Vocabulary,OOV)数量更多,导致目前维汉机器翻译的结果中有很多 UNK(遇到 OOV,一般的处理方法是在译文中用 UNK 表示),所以对于维吾尔语口语用语的词项归一化很有必要[1-2]。

如:“询问过程中”一词对应的维吾尔语为 تەكشۈرۈلى ۋې تېپىتۇ ,但是口语文本中经常使用 سۈرۈشتۈرۈل ۋې تېپىتۇ ,而 سۈرۈشتۈرۈل ۋۇې تېپىتۇ 无法在目前受限规模的维汉双语平行口语语料训练的机器翻译系统中被正确翻译出来,但是如果将它归一化到正规书写的方式,句子能在意思未改变的前提下被正常翻译出来,这也是本文的最终目的。

词汇归一化作为语料预处理的一个关键步骤,一直以来吸引了很多研究者的目光。最早的也是最简单的可以用于词汇归一化的方法便是噪声信道模型[3],对于非正规语料 T 与之对应的正规语料 S,这个模型包含 2 个部分:语言模型 $P(S)$ 和一个归一化模型 $P(T|S)$。如果将非正规用语的文本当作语言 T,它对应的正规文本作为 S,根据 $P(S|T)=P(T|S)\times P(S)/P(T)$,$P(T)$ 是固定的,那么通过求解 $\arg\max P(T|S)\times P(S)$ 来求解对应的 S,从而求到 $\arg\max P(S|T)$,S 便是 T 归一化后的结果。文献[4]将噪声信道模型运用到归一化中,之后对噪声信道模型进行扩展[5],将词的发音作为特征加入模型中。但是这种模型都是有监督的模型,需要大量的标注语料对模型进行训练。文献[6]对噪声信道模型进行无监督训练扩展。

另一个比较主流的词归一化方法是基于统计机器翻译的方法。文献[7]提出一个编码/解码为字符级别的短语统计机器翻译系统,使用非正规书写的英语为源语言,对应的正规书写的英语为目标语言,通过大量语料训练出来的这个翻译系统能很好地处理归一化问题。跟噪声信道模型类似,训练阶段需要大量的训练数据,但是一一对应的非正规和正规的语料是很难大规模获取的,对于维吾尔语这种语料匮乏的小语种难度更大。

近年来,类似于基于上下文的图的无监督的随机游走[8]算法用于社交媒体上的文本的归一化,之后研究热点已经转向无监督的方法。文献[9]把 2 个词的上下文的相关性当作 2 个词的相关性的依据,从而用来做归一化。文献[10]使用类似于文献[9]的方法,利用深度神经网络和 word2vec 进行未登录词与词典内的正规词进行相似度比较,最后使用语言模型来筛选。本文提出的方法与文献[10]方法类似,将其提出的方法引入到维吾尔语的归一化中,但是考虑由于一句话中可能包含多个非规则化的词语,一次归一化过程并不能完全将非正规词归一化,从而在文献[10]方法的基础上引入 bootstrapping[11]重采样策略[12-13],每一遍归一化之后,重新采样,迭代直到非规则化的词替换次数未达到某个阈值停止;而且本文方法只是用于词,由于维吾尔语的短语划分不稳定,短语级别的归一化将作为以后的研究方向。

2　维吾尔语无监督词汇归一化模型

本文使用贪心解码器以及引入 bootstrapping 策略得到维吾尔语无监督词汇归一化模型,模型如图 1 所示。首先是对网络爬取的非正规用语语料进行初步的数据预处理,预处理操作只是最简单的筛选,将一半以上单词都是非正规词的句子剔除,这主要是为了保证解码过程的正确性;然后将正规用语语料和非正规用语语料放在一起,训练出词向量,再根据从正规用语的语料中抽取的正规用语词典,在向量空间中找到每个非正规词的 k 近邻当作候选表,之后使用贪心解码器对非规则化文本中每个句子依据语言模型和字符串相似度选择非规则化词的最优解,遍历完了将替换之后的非正规用语文本跟正规用语文本一起重新训练词向量,一直递归执行直到满足退出条件。

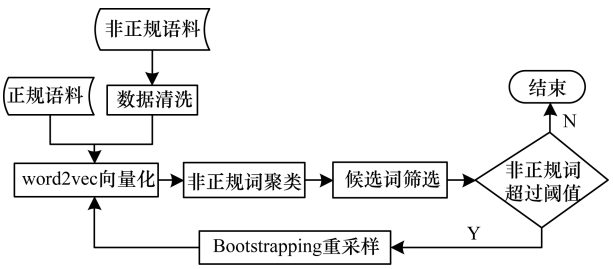

图 1 无监督词项归一化模型

核心算法流程的伪代码为：

输入 正规用语语料库 $StdS=\{s_1,s_2,\cdots,s_n\}$，非正规用语语料库 $UStd=\{s_1,s_2,\cdots,s_n\}$，评分阈值 *threhold*

matchpair = {}

while 匹配量大于阈值 do

UStd 句子进行清洗

StdS 与 UStd 一起训练出 word2vec 模型 model

StdS 训练出语言模型 langModel

StdS 抽取生成正规词字典 NormalDict，UStd 抽取生成非正规词字典 UnNormalDict

for each word in UnNormalDict do

根据 model 找出 cosine 相似度最大的 topn 的 candidates

for each line in UStd do

根据 langModel 计算 line 的困惑度 perp1

计算 line 将 word 替换为 candi 之后句子的困惑度 perp2

根据 ratio 和字符串相似度的综合评分 score 重排序 candidates

for each word in UnNormalDict do

从其 candidates 里面找到符合条件的匹配，加入到 matchpair 中

依据 matchpair 替换 UStd 中匹配到的非正规词迭代

end

2.1 词向量

对词进行向量化表示一直是热点，从最初的空间向量模型，到浅层语义分析（Latent Semantic Analysis）、PCA 等，但这些向量都是基于词共现来实现，并不能把握住语义信息。百度提出神经网络搭建二元语言模型的方法[14]，文献[15]提出了基于神经网络的语言模型之后，后续涌现出一批使用神经网络生成词向量的方法，比较具有代表性的有 google 提出的 word2vec[16] 和 glove[17]。使用神经网络训练出来的词向量，考虑到了上下文信息，所以对词意的表现力比之前的向量表示更加强[18]。对于词汇的归一化便可以考虑使用词向量作为一个特征，因为那些拼错了或者不同形式的词，它们的上下文还是比较相似的。

传统的将词向量化的方法都是将词用一个 one-hot 的向量表示，但是这种方法遇到的问题就是数据的稀疏，而且向量除了表示词以外，并没有将词的上下文、语义上表达出来。词的分布式表示提出之后，由于这种向量能很好地表达出词之间的相似性，很快为研究者所青睐。通过训练将每个词映射成 K 维实数向量（K 一般为模型中的超参数），通过词之间的距离（比如 cosine 相似度、欧氏距离等）来判断它们之间的语义相似度。

词的分布式表示是神经网络语言模型的代产物，神经网络语言思路与 n-gram 模型类似，使用 $w_{t-n+1},w_{t-n+2},\cdots,w_{t-1}$ 来预测 w_t，$\boldsymbol{C}(w)$ 是词 w 对应的词向量，神经网络语言模型使用一套唯一的词向量，存在矩阵 $\boldsymbol{C}$ 中，$\boldsymbol{C}$ 的大小为 $|V|\times m$，$|V|$ 是词表大小，m 是向量的维度，从词 w 到 $\boldsymbol{C}(w)$ 就是从矩阵 $\boldsymbol{C}$ 中取出对应的那一行。此模型如图 2 所示，是一个三层的神经网络，网络的第 1 层是将窗口中的词对应的词向量 $\boldsymbol{C}(w_{t-n+1}),\cdots,\boldsymbol{C}(w_{t-2}),\boldsymbol{C}(w_{t-1})$ 拼接起来，形成一个 $(n-1)m$ 的向量，记为输入 x。

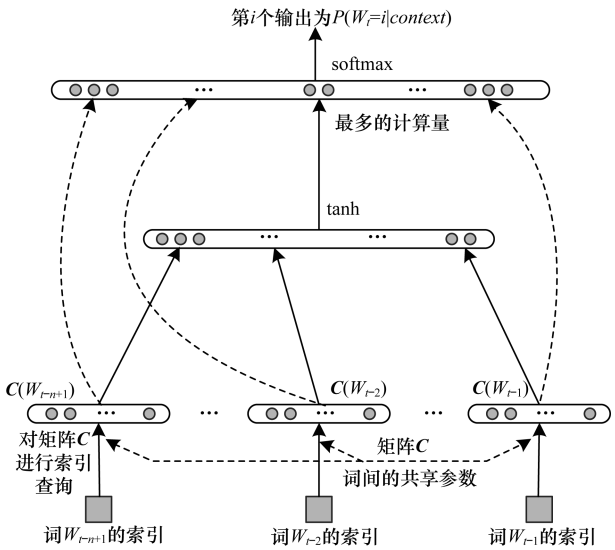

图 2 神经网络语言模型

网络的第 2 层就是对输入进行一个非线性变换：

$$h=\tanh(d+Hx) \tag{1}$$

网络的输出层一共有 $|V|$ 个节点，每个节点 y_i 表示下一个词为 i 的未归一化 log 概率。最后使用 softmax 将输出值归一化成概率。

$$y=b+Wx+U\tanh(d+Hx) \tag{2}$$

此模型的目标函数如式(3)所示，通过最大化下一个词的概率的训练过程，矩阵 $\boldsymbol{C}$ 作为参数的一部分进行梯度下降调优，最后这个矩阵便是词向量。这样训练出来的词向量具有很好的语义表示能力。

$$L=\frac{1}{T}\sum_t \log_a f(w_t,w_{t-1},\cdots,w_{t-n+1};\theta)+R(\theta) \tag{3}$$

实验采用的是 Word2Vec 工具生成的词向量。Word2Vec 有 2 种方式：CBOW 和 skip-gram，采用 skip-gram、skip-gram 的目的是使用一个词来预测窗口内的其他词，最大化其他词的概率。

由于通过 word2vec 可以将单词投射到低维向量空间，本文采用2 个词的向量的 cosine 距离作为 2 个词的相似度，用于候选词的初级筛选，如图 3 所示。

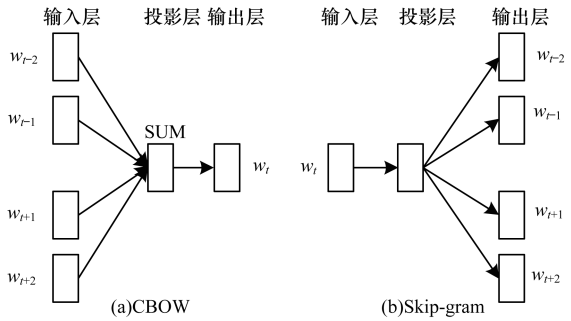

图 3　word2vec 的 2 种方式

2 个维度为 D 的向量 $\boldsymbol{e}$ 和 $\boldsymbol{f}$ 的 cosine 距离定义如下:

$$cosine\_distance(\boldsymbol{e},\boldsymbol{f}) = \frac{\sum_{i=1}^{D}\boldsymbol{e}_i \times \boldsymbol{f}_i}{\sqrt{\sum_{i=1}^{D}(\boldsymbol{e}_i)^2 \times \sum_{i=1}^{D}(\boldsymbol{f}_i)^2}} \quad (4)$$

2.2　贪心解码算法

在非正规词聚类之后,每个非正规词都有一个候选正规词表,从这个词表中选出该词意思最近的正规词作为此非正规词的归一化目标。对于包含非正规词的句子,从候选词表中选择最优解可以类比为一个简易的机器翻译的解码过程,只需要针对非正规词进行部分解码即可。

采用一个比较简单的贪心策略的解码器进行候选词的筛选,贪心策略的评分价值采用句子的语言模型困惑度评分变化率以及非正规词与其候选词的字符串相似度的综合考虑。选取得分超过阈值的词。评分如式(5)所示,*pp_ratio* 是语言模型困惑度变化率,*lexicalsimilarity* 是 2 个词的词汇字符串相似度,λ_1、λ_2 分别是模型的 2 个超参数,在实验中使用手工调优得到,手动调参策略是固定一个 threshold,将 λ_1、λ_2 均以 0.5 为初值,学习率为 0.03,以正则化之后测试集的 BLEU 值作为评价指标,从而选取较优的比例;threshold 的选取则是在 λ_1、λ_2 选取之后手动进行调整,策略与 λ_1、λ_2 调参类似。

$$score(word) = \lambda_1 pp\_ratio + \lambda_2 lexicalsimilarity \quad (5)$$

语言模型的目的是建立一个能够描述给定词序列在语言中出现的概率的分布,使用一个采用 Kneser-Nye 平滑的 $n=5$ 的 n-gram 语言模型对句子进行困惑度打分。困惑度评分如式(6)所示。

$$PP(S) = \sqrt[N]{\prod_{i=1}^{N}\frac{1}{P(w_i \mid w_{i-4}w_{i-3}w_{i-2}w_{i-1})}} \quad (6)$$

对于候选词的困惑度 *pp_ratio* 打分为用此候选词替换对应的非正规词之后的句子的困惑度变化率,计算方式如式(7)所示。

$$pp\_ratio = \frac{PP(S) - PP(\text{替换后的句子 } S)}{PP(S)} \quad (7)$$

对于字符串的相似度最常被采用的是编辑距离,编辑距离又叫 Levenstein 距离,是指 2 个字串之间,由一个转成另一个所需的最少编辑操作次数。许可的编辑操作包括将一个字符替换成另一个字符,插入一个字符,删除一个字符。一般来说,编辑距离越小,2 个串的相似度越大。

但是编辑距离并不能特别适合这种场景,因为一般非正规书写的单词包括很大一部分是对单词进行大面积的缩写。采用文献[6]所提出来的词汇相似度值,2 个单词 S_1、S_2 的词汇字符串相似度如式(8)所示,是 2 个字符串的最长公共子串率与编辑距离之除,这个相似度很好地适用于缩写的情况。

$$lexicalsimilarity(S_1,S_2) = \frac{LCSR(S_1,S_2)}{ED(S_1,S_2)} \quad (8)$$

2 个单词的最长公共子串率如式(9)所示,是 2 个字符串的最长公共子串与它们的最长长度之除。

$$LCSR(S_1,S_2) = \frac{LCS(S_1,S_2)}{maxlength(S_1,S_2)} \quad (9)$$

解码算法的伪代码如下:

输入　非正规用语语料库 $UStd = \{s_1, s_2, \cdots, s_n\}$,评分阈值 *threshold*

```
matchpairs = {}
for sentence in UStd do:
    计算 sentence 的语言模型评分 pp1
    for 非正规词 UFword in sentence do:
        计算候选集中一个正规词 FWord 替换后的句子语言模型评分 PP2
        计算 score(FWord)
    if max(score(FWord)) > threshold do:
        将 UFword 替换为 FWord 继续当前句子解码
    else do:
        进行下一句子解码
end
```

2.3　bootstrapping

解码器使用的都是基于很多噪声的语料训练出来的向量空间以及上下文信息,会导致一些非正规词对应的正规词不能聚类到 top-k 的候选集中,从而不能在解码中匹配出来。为了解决这个问题,引入了 bootstrapping 方法。bootstrapping 是统计学中的重采样,本文应用 bootstrapping 是带更新的重采样,在所有句子解码完成后,将匹配到的非正规词归一化为其对应的正规词,将修改过的语料与正规语料一起,再重采样,进行递归来对之前归一化过程中未能归一化的词进行进一步的归一化。重采样的策略采用.632 自助法,对于包含 d 个样本的数据集,有放回地抽样 d 次,产生 d 个样本的数据集,每次递归之后采取的重采样操作一样。

3　实验结果与分析

本文主要研究对象为维吾尔语口语中非正规词,首先实验验证词归一化模型的准确性,然后将归

一化的结果运用于维汉机器翻译中验证本文方法对机器翻译系统的作用的有效性,作为对比对象,引入了文献[10]提出的方法。

3.1 正确性分析

由于尚未有通用的维吾尔语词归一化方法,采用人工判定方式,使用的语料资源为:0.2 MB 的新闻维汉双语语料作为正规语料,0.2 MB 的网络文本作为非正规语料;使用的词向量是 word2vec,窗口大小为 8,最小出现次数为 10 而训练生成的 200 维的向量,语言模型为使用 kenlm[19] 对此正规语料训练的 $N=5$ 的 N-gram语言模型。

使用本文方法能成功归一化 1 812 次非正规词,对归一化成功的词进行准确度、召回度和 F1 值评价,结果如表 1 所示。

表 1 归一化词正确性分析 %

方法	准确度	召回度	F1 值
基线	50.6	61.1	55.4
文献[10]方法	42.5	61.3	50.2
本文方法递归 1 次	48.9	59.9	53.8
本文方法递归 2 次	51.7	58.0	54.7
本文方法递归 3 次	52.4	57.1	54.6
本文方法递归 4 次	53.0	56.2	54.6

从实验结果可以看出,本文方法与文献[10]方法均能够在此场景中有效地进行词的归一化。本文方法在递归的进行中,准确度逐步增加,并且在递归 3 次之后就优于文献[10]方法,这说明本文方法引入的 bootstrapping 策略能有效地提高归一化的准确性。在召回率上,本文方法随着递归的进行,召回率逐步降低,主要是由于前序递归中为正确归一化的词引入的噪声,最后召回率低于文献[10]方法,但总体 F1 值也与文献[10]方法相当。

3.2 机器翻译实验验证

把本文提出的非正规词归一化方法应用于实际的机器翻译系统中,来验证该方法的有效性。

归一化模块的实验设置:使用的语料资源为 0.2 MB的新闻维汉双语语料作为正规语料,0.2 MB的网络文本作为非正规语料,使用的词向量是 word2vec 生成的窗口大小为 8 的 200 维的向量,语言模型为使用此正规语料训练的 $N=5$ 的 N-gram 语言模型,超参数 λ_1、λ_2 经过多次实验,采用 0.43、0.57 效果最优。

机器翻译实验设置:采用维汉新闻语料和未正规化的口语语料作为实验对象,训练集采用CWMT2015 的维汉新闻语料,由于尚未有公开的维汉双语口语语料集,实验采用爬取以及标注的网页论坛语料作为测试集,语料样本规模如表 2 所示。

表 2 机器翻译语料

训练集	开发集	测试集 1	测试集 2
0.2 MB	20 KB	1 KB	1 KB

实验的基线系统为 moses3.0[20] 训练的基于短语的统计机器翻译系统[21],该系统基于最小错误率训练方法优化翻译系统权重,最后采用 BLEU 值作为评价指标。本文设置如下 3 个翻译实验:

1)基线:利用新闻语料训练的统计翻译模型直接对测试集进行翻译。

2)文献[10]方法:对测试集使用文献[10]方法进行归一化之后利用基线进行翻译。

3)本文方法:采用本文提出的归一化方法进行归一化之后利用基线进行翻译,递归 i 表示本文方法进行重采样递归 i 次之后进行归一化的结果。

实验结果如表 3 所示,利用本文方法进行归一化之后的文档的翻译结果的 BLEU 值有了显著的提升。在递归 2 次之后,本文方法的结果略优于进行一次解码的文献[10]提出的方法。

表 3 机器翻译实验结果

系统	测试集 1	测试集 2	平均值
基线	18.46	17.36	17.91
文献[10]方法	19.02	17.83	18.43
本文方法递归 1 次	18.85	17.70	18.27
本文方法递归 2 次	19.09	17.93	18.51
本文方法递归 3 次	19.15	18.06	18.60
本文方法递归 4 次	19.20	18.09	18.64

本文方法的有效性在于使用了无监督的归一化方法在对测试集进行前编辑之后,能将测试集中不能被有限的平行语料训练出来的统计机器翻译系统中的一些非正规词替换为能正确翻译出来的书写正确或意思相近的词,从而翻译的效果有了提升。如图 4 所示,测试语料中 سۈرۈشتۈرۈل ۋۋېتىپتۇ 这一词并不能使用当前平行语料进行翻译,对于包含这个词的句子,机器翻译系统输出的译文有很多未登录词。但是使用本文方法能将本次归一化为 تەكشۈرۈلى ۋې تىپتۇ (在调查中),句子通过本文方法进行归一化之后,机器翻译的译文效果有了显著的提升,如图 5 所示。

```
Translating: ن ۋۇتە ، شوپۇرنىڭ ئىز - دېرىكى بولمىغان بولۇپ ، ۋەقەنىڭ يۈز بېرىش س ەبى سۈرۈشتۈرۈل ۋۋېتىپتۇ .
Line 14: Initialize search took 0.000 seconds total
Line 14: Collecting options took 0.004 seconds at moses/Manager.cpp Line 141
Line 14: Search took 0.432 seconds
目前 ， 开车 的 人 ， 失踪 的 ， 事故 发生 的 原因 是 سۈرۈشتۈرۈل ۋۋېتىپتۇ 。
```

图 4 包含非正规词的句子实例

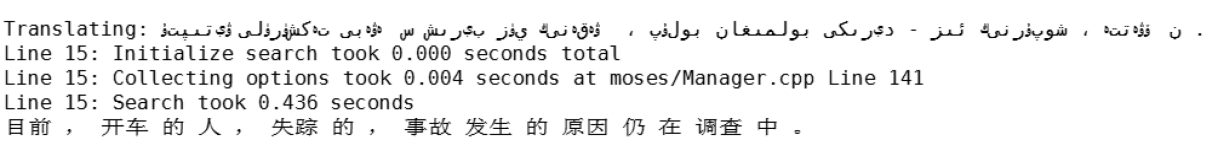

图 5　归一化之后的句子实例

本文方法的效果随着递归的进行,BLEU 值逐步趋于收敛,这是由于本文方法每轮递归中未正确归一化的词所引入的噪声导致后续的递归过程中能进行正确归一化的词数量减少所导致的。

4　结束语

本文提出了一种无监督的维吾尔语口语中非正规词的归一化方法,将该方法运用于维汉机器翻译的待翻译句子的前编辑归一化之后,相比于基线系统,使用不同领域训练的统计机器翻译系统,在测试集上 BLEU 值提升了 0.7。此外本文方法也是对文献[10]方法的一种改进,引入了 bootstrapping 方法并且采用了另一个解码器以及不同的打分机制,实验结果也证明本文方法有一定的改进,在准确度上有了 2.4 个百分点的提高,由于引入重采样策略,召回率降低了 5 个百分点,在机器翻译上,本文方法也较之在 BLEU 值上提高了 0.2。

由于本文未能引入更多的维吾尔语的语言学特性,因此后续将在解码器中加入部分语言学方面的规则,进一步提高归一化的召回率。

参考文献

[1] 年　梅,张兰芳. 维吾尔文网络查询扩展词的构建研究[J]. 计算机工程,2015,41(4):187-189,194.

[2] MI Chenggang, YANG Yating, ZHOU Xi, 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.

[3] SHANNON C E. Communication Theory of Secrecy Systems[J]. Bell System Technical Journal, 1949, 28(4): 656-715.

[4] 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. [S. l.]: Association for Computational Linguistics, 2000: 286-293.

[5] TOUTANOVA K, MOORE R C. Pronunciation Modeling for Improved Spelling Correction [C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2002: 144-151.

[6] COOK P, STEVENSON S. An Unsupervised Model for Text Message Normalization [C]//Proceedings of Workshop on Computational Approaches to Linguistic Creativity. [S. l.]: Association for Computational Linguistics, 2009: 71-78.

[7] AW A T, ZHANG Min, XIAO Juan, et al. A Phrase-based Statistical Model for SMS Text Normalization [C]//Proceedings of COLING/ACL on Main Conference Poster Sessions. [S. l.]: Association for Computational Linguistics, 2006: 33-40.

[8] HASSAN H, MENEZES A. Social Text Normalization Using Contextual Graph Random Walks[C]//Proceedings of the 51st Annual Meeting Computational Linguistics Meeting. [S. l.]: Association for Computational Linguistics, 2013: 1577-1586.

[9] HAN Bo, COOK P, BALDWIN T. Automatically Constructing a Normalisation Dictionary for Microblogs[C]//Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. [S. l.]: Association for Computational Linguistics, 2012: 421-432.

[10] SRIDHAR V K R. Unsupervised Text Normalization Using Distributed Representations of Words and Phrases[C]//Proceedings of Workshop on Vector Space Modeling for Natural Language Processing. New York, USA: ACM Press, 2015: 8-16

[11] MOONEY C Z, DUVAL R D, DUVAL R. Bootstrapping: A Nonparametric Approach to Statistical Inference [J]. Technometrics, 1993, 36 (4): 435-436

[12] 罗　军,高　琦,王　翊. 基于 Bootstrapping 的本体标注方法[J]. 计算机工程,2010,36(23):85-87.

[13] 何婷婷,徐　超,李　晶,等. 基于种子自扩展的命名实体关系抽取方法[J]. 计算机工程,2006,32(21):183-184.

[14] XU W, RUDNICKY A I. Can Artificial Neural Networks Learn Language Models? [D]. Pittsburgh, USA: Carnegie Mellon University, 2000.

[15] BENGIO Y, DUCHARME R, VINCENT P, et al. A Neural Probabilistic Language Model [J]. Journal of Machine Learning Research, 2003, 3(2): 1137-1155.

[16] MIKOLOV T, KARAFIÁT M, BURGET L, et al. Recurrent Neural Network Based Language Model[C]//Proceedings of Conference of the International Speech Communication Association. Berlin, Germany: Springer, 2010: 1045-1048.

[17] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global Vectors for Word Representation[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Berlin, Germany: Springer, 2014: 1532-1543.

[18] 张为泰. 基于词向量模型特征空间优化的同义词扩展研究与应用[D]. 北京:北京邮电大学,2014.

[19] HEAFIELD K. KEN L M: Faster and Smaller Language Model Queries[C]//Proceedings of the 6th Workshop on Statistical Machine Translation. [S. l.]: Association for Computational Linguistics, 2011: 187-197.

[20] KOEHN P, HOANG H, BIRCH A, et al. Moses: Open Source Toolkit for Statistical Machine Translation[C]//Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. [S. l.]: Association for Computational Linguistics, 2007: 177-180.

[21] CHIANG D. Hierarchical Phrase-based Translation [M]. [S. l.]: Association for Computational Linguistics, 2007.

编辑　顾逸斐