文章编号：1003-0077（2011）00-0000-00

基于稳健词素序列和LSTM的维吾尔短文本分类研究*

沙尔旦尔·帕尔哈提，米吉提·阿不里米提，艾斯卡尔·艾木都拉*

（新疆大学信息科学与工程学院，新疆 乌鲁木齐 830046）

摘要：本文讨论了基于 Word2vec 和长短期记忆（LSTM）网络的维吾尔短文本分类技术。使用基于词-词素平行语料的稳健词素切分和词干提取方法，从噪声文本中提取词干后分别建立词和词干集合，并通过 word2vec 工具映射到实数向量空间。然后采用 LSTM 网络作为特征选择和文本分类算法进行维吾尔短文本分类实验。结果显示，在基于词干向量的分类实验中得到 95.48%的分类准确度。从实验结果看，对于派生类语言而言，特别是对于带噪声的文本，基于词干的分类方法有更多优异性能。

关键词：维吾尔语；文本分类；LSTM 网络；形态学

中图分类号：TP391　　　　　**文献标识码：**A

Uyghur short text classification based on word stem and LSTM network

Sardar Parhat, Mijit Ablimit, Askar Hamdulla*

(Information Science and Engineering Institute, Xinjiang University, Urumqi, Xinjiang, 830046,China)

Abstract: This paper discusses Uyghur short text classification technique based on Word2vec and LSTM network. The robust morpheme segmentation and stemming extraction methods based on word-morpheme parallel corpora are used to extract the stems from noisy texts. The set of stems and words are established respectively. The stems are mapped to real vector space by Word2vec tool. Then we use the LSTM network as feature selection and text classification algorithm to implement Uyghur text classification experiments. The results show that 95.48% of classification accuracy is obtained in the stem-vector based classification experiment. From the experimental results it can be seen that for derivative languages, especially for noisy texts, stem-based classification method has more excellent performance.

Key words: Uyghur; text classification; LSTM network; morphology

---

* **收稿日期：**　　　　　　**定稿日期：**

基金项目：国家自然科学基金项目（61662078，61633013）

作者简介：沙尔旦尔·帕尔哈提（1984——），男，博士研究生，主要研究方向为文本及图像信息检索；米吉提·阿不里米提（1974——），男，副教授，硕士生导师，主要研究方向为语音和语言信息处理；艾斯卡尔·艾木都拉（1972——），男，教授，博士生导师，主要研究方向为图像处理与模式识别、智能信息检索。

1 引言

近年来，随着互联网的快速发展，每天都会产生大量的文本、音频、图片和视频数据，其中文本信息的数据量最大。但文本信息混乱，难以人工区分和组织。因此，对文本数据进行自动分类已经成为一项紧迫的工作。

维吾尔语的句子由自然分开的词组成。词由词干追加词缀来派生，因此词汇量巨大。词缀提供语义及语法功能，如表 1 所示。

表 1 维吾尔词语变体

词干	变体	词缀
（工）vix	（工人）vixci =vix+ci	ci
	（办公室）vixHana = vix+Hana	Hana
	（职位）vixtat = vix + tat	tat

以上表拉丁文对应的维吾尔字母对照表如表 2 所示：

表 2 维吾尔与拉丁字母对照表

序号	拉丁	维吾尔	序号	拉丁	维吾尔	序号	拉丁	维吾尔
1	y	ي	12	p	پ	23	e	ې
2	a	ا	13	m	م	24	q	ق
3	l	ل	14	s	س	25	H	خ
4	G	غ	15	b	ب	26	U	ۈ
5	u	ۇ	16	d	د	27	h	ھ
6	z	ز	17	A	ە	28	g	گ
7	k	ك	18	v	ئ	29	f	ف
8	x	ش	19	r	ر	30	w	ۋ
9	i	ى	20	n	ن	31	O	ۆ
10	t	ت	21	N	ڭ	32	J	ژ
11	o	و	22	c	چ	33	j	ج

维吾尔文字中的 Amza 符号我们用拉丁文字母 v 来表示。

维吾尔词语中，词干是具有实际意义的词汇单元。词干提取能够使我们抓有效的、有意义的特征，并大大减少特征的重复出现率和特征位数，如以下例子所示：

vixci vixHana vixini vixlAp tUgAtty.

以上句子词素切分后变成以下所示：

vix+ci vix+Hana vix+ini vix+lAp tUgAtty.

以上句子中有 5 个词语，其中前四个词语的词干都是 vix（工），这样，一个词干能够抓四个词特征，特征位数会大幅减少。

维吾尔语自然语言处理（NLP）的主要问题是资源的缺乏和形态结构的变化，从因特网上收集的数据在编码和拼写等方面有带噪声和不确定性等特点[1]。方言和在拼写和编码方面的不确定性对提取和分类短文本和带有噪声的文本数据的可靠性提出了巨大挑战[2]。然而，提取和分类短、有噪声的文本数据是维吾尔语自然语言处理不可避免的重要步骤。

部分学者发表了维吾尔语词干提取有关的研究结果[3-4]。文[3]根据简单的名词构形词缀规则进行构词成分有限状态分析，以此提取维吾尔语的词干。文[4]根据维吾尔语的构词约束条件，用词性特征和上下文词干信息来提取维吾尔语词干。文[4]没有考虑句子级别的上下文信息。以前的这些词干提取有关的工作大多是基于简单的后缀为基础的词干方法和一些简单的人工设置的规则。因此存在歧义，尤其是在短文本上。基于句子或较长上下文的可靠的词干能够正确地预测噪声环境中的词干和词，为自然语言处理的许多方面提供了有效的途径。我们的多语言文本处理工具[2]可以为整个句子提供形态分析，并减少噪声文本中的歧义。有些学者对维吾尔语文本分类做了一些研究[5-7]。文[5]以 KNN 为分类器对维吾尔语文本进行分类实验。在本研究中用词频-逆文档频率（TFIDF）算法，来计算特征的权重值。文[6]利用 TextRank 算法对维吾尔语文本进行句子情感分类，并选用 SVM 作为分类算法。文[7]用词频，对传统的 TFIDF 权值函数进行加权来改进，用贝叶斯分类器进行了维吾尔文本分类实验。

自然语言具有结构依赖性的特点。以往研究者在维吾尔文本分类中所使用的方法是在传统的分类框架下进行的，这些方法对文本中单词的频率和一些子词单元进行简单的统计，其中所使用的机器学习过程较浅，不考虑文本中词语之间的语义关系，因此无法保持文本上下文的清晰语义信息。

自动文本分类是一个引导性的学习过

程，其中大量的非结构化文本信息（文本文档、网页等）根据给定的分类系统和文本信息的内容自动分类为指定的类别[8-9]。自动文本分类在许多信息检索工作和相关任务中得到广泛应用，包括情感分析[10]、垃圾邮件过滤[11]和网页搜索[12]。

长短期记忆网络（LSTM）具有捕获顺序数据之间的依赖关系的能力，因为它善于捕获长距离信息，所以在语音识别中取得了显著的准确性[13-15]。LSTM 网络可以有效的解决当用传统的分类模型进行自动文本分类时文本中词语上下文意义被忽视的问题。

文本表示和特征选择是文本挖掘和信息检索中的基本问题。它量化从文本中提取的特征词来表示文本信息。词袋模型（BOW）[16]和 TFIDF[17]常用于表示文本特征。本文提出基于词和词干向量的维吾尔语文本分类方法。采用基于形态规则的词素和词干切分方法稳健地提取其词干，以 word2vec 算法更好地表示维吾尔文本及其词和词干特征，并利用 LSTM 网络作为特征选择和文本分类算法，得到维吾尔语文本分类模型。我们从互联网收集文本，以建立文本语料库，使用 LSTM 网络自动选择维吾尔语文本特征，并对此语料库进行分类实验。

2 提出的维吾尔文本表示和分类方法

我们的分类算法主要包括两部分。一是维吾尔语文本集的预处理，包括对实验文本数据的获取、对词干的降噪。二是分类过程，包括特征提取和分类。

2.1 词向量文本表示法

近期，深度神经网络和表示学习[18-19]提供了更好的文本表示方法和缓解数据稀疏问题。Mikolov 等人[20]提出了 word2vec 文本表示方法，并利用深度学习和向量运算的思想，通过训练把文本内容的处理简化到 N 维向量空间，寻求文本数据更深层次的特征表示，并使用在向量空间中相似性来表示文本的语义相似度。

词向量是一个真值向量[21]，通过计算任意两个给定词向量或词干向量之间的距离，来可以容易地找到词或词干相似度。我们利用 word2vec 算法可以快速有效地训练词和词干向量。Word2vec 算法包括两个重要的子模型：CBOW（连续词袋）模型[22]和 Skip-gram 模型[23]。

CBOW 是一个在给定上下文词 $W_{t-c},W_{(t-c)-1},\cdots,W_{t-1},W_{t+1},W_{t+2},\cdots,W_{t+c}$ 的条件下预测特定单词 W_t 发生的概率 $P(W_t|W_{t-c},W_{(t-c)-1},\cdots,W_{t-1},W_{t+1},W_{t+2},\cdots,W_{t+c})$ 的模型，如图 1 所示。在这个模型中，一个词由在这个词前后的 c 个词表示，c 是预选窗口的大小，输出是这个特征词 W_t 的词向量，如图 1 所示。我们将使用 CBOW 特征表示模型从噪声表达式中得到词干向量。

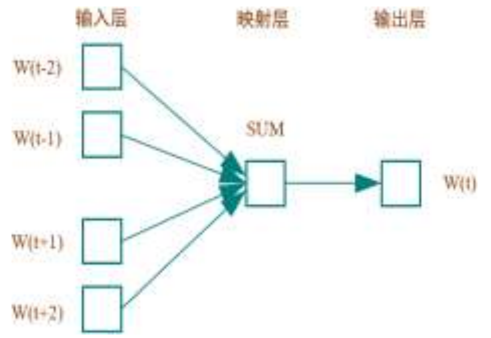

图 1 CBOW 模型

Skip-gram 模型的思想正好与 CBOW 模型相反，即，它在给定特定词 W_t 的条件下，预测上下文词 $W_{t-c},W_{(t-c)-1},\cdots,W_{t-1},W_{t+1},W_{t+2},\cdots,W_{t+c}$ 的发生概率 $P(W_{t-c},W_{(t-c)-1},\cdots,W_{t-1},W_{t+1},W_{t+2},\cdots,W_{t+c}|W_t)$。如图 2 所示。

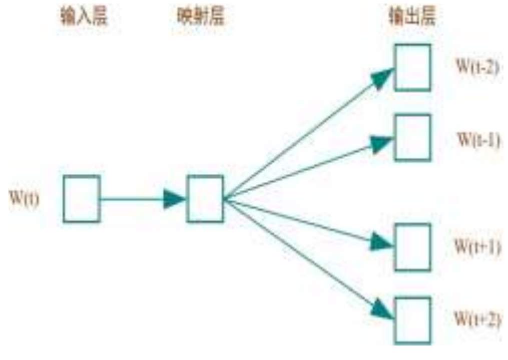

图 2 Skip-gram 模型

通过 word2vec 训练得到的词（词干）向量可以通过其余弦距离来判断语义相似程度。计算得到的余弦值越大，语义越相近；反之，语义相差越远，如表 3 所示。

表 3 词向量语义相似度

词 muzika(音乐) 相关词		词 tor(网络) 相关词	
词	余弦距离	词	余弦距离
keciliki(晚会)	0.8218	bekiti(网站)	0.9206
vusul(舞蹈)	0.8138	simsiz(无线)	0.8783
sAnvAt(艺术)	0.7970	kOcmA(移动)	0.8704
naHxa(歌曲)	0.7742	sodisi(商务)	0.8694
gitar(吉他)	0.7413	vUndidar(微信)	0.8664

表 3 中可以看出分别输入词 muzika（音乐）和 tor（网络），并通过计算词向量之间的余弦距离，来得到的与这两个输入词语义最相近的 5 个词。

2.2 LSTM 网络框架

LSTM 网络是一种时间循环的神经网络，适用于处理和预测时间序列中间隔较长和延迟较大的重要事件[24]。LSTM 网络的一般架构如图 3 所示。每个节点包含三个门结构，分别为遗忘门、输入门和输出门。

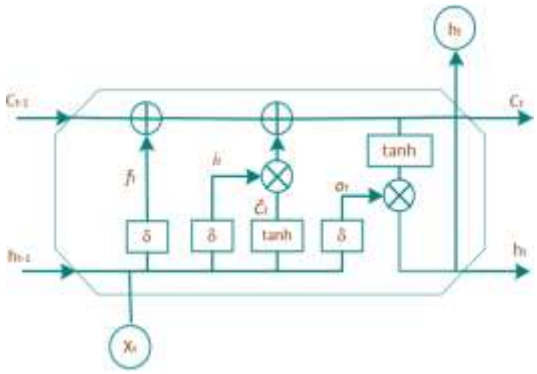

图 3 LSTM 网络结构

在一个计算中，LSTM 节点首先计算遗忘门，公式如 2-1 所示。

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_t) \qquad 2\text{-}1$$

其中，W_f 和 x_t 是遗忘门中 *Sigmoid* 层的参数矩阵和偏置矩阵， f_t 是遗忘向量，其中每一位都对应于 C_{t-1} 中的一个数据，用于控制信息的流入。

第二步是将输入数据 x_t 存储到节点中，并更新节点的状态。该步骤首先计算出第二个门结构，即输入门的 *Sigmoid* 层与一个 tanh 层，并得到两个选后值 i_t 和 $\overline{C}_t$，如公式 2-2 和 2-3 所示。

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i) \qquad 2\text{-}2$$

$$\overline{C}_t=\tanh(W_c\cdot[h_{t-1},x_t]+b_c) \qquad 2\text{-}3$$

然后，根据公式 2-4，得到节点的新状态 C_t。

$$C_t=f_t*C_{t-1}+i_t*\overline{C}_t \qquad 2\text{-}4$$

最后，节点通过第三门结构，即输出门，获得输出向量，如公式 2-5 所示。

$$o_t=\sigma(W_0\cdot[h_{t-1},x_t]+b_0)$$

$$h_t=o_t*\tanh(C_t) \qquad 2\text{-}5$$

通过增加门结构，LSTM 能够有效地避免了传统的 RNN 模型训练困难这一问题，并有效的弥补了在传统 RNN 网络上执行反向传播算法时梯度爆炸或者梯度消失等缺点。

2.3 稳健的维维吾尔语词素切分

由广泛的跨语言和跨文化交流所引起的不确定拼写的嘈杂文本导致不断创造的新词，新概念和新表达。这些新词大多是新近借用的外来词或词干，以及由于拼写习惯的不同和方言的变形而引起的噪音整合。书写形式不确定性的另一个原因是书写系统的历史变化。例如，维吾尔语目前使用阿拉伯字母，但 30 年前使用了罗马字母。在更古老的时代，甚至更多的书写形式被使用。这些不同的书面系统在现代社会留下了它们的遗产，虽然不太可能在官方媒体上出现，但在网上论坛和聊天工具无处不在。

我们开发了一个紧凑的工具以改进少数民族语言自然语言处理[2]。该工具将单词序列切分成粘着性语言的词素序列。该工具在功能和语言上都是可扩展的。

该工具根据对齐的词语和词素并行训练数据，从训练数据中自动学习各种表面形式和声学规律。当词素被合并为一个词时，边界上的音素根据语音和谐规则改变其表面形式。语音将互相和谐，互相吸引对方的

发音。当发音准确地表示时，可以在文本中清楚地观察到语音和谐。切分程序将导出每个候选的所有可能的切分形式。一个独立的统计模型可以被纳入，从 N 个好的结果中选择最佳结果。该工具包为词干提取提供了可靠的依据，极大地改进了短文本分类任务。

我们使用词-词素平行语料库训练统计模型，如图 4 所示。

图 4 词素切分流程

该工具在包括 10025 个句子的 64.51k 大小的词-词素平行语料库上训练统计模型，其中选择了 9025 个句子为训练语料库，1000 个句子为测试语料库，做了词素切分和词干提取实验，其词干提取准确度最高时达到 97.66%。这是所有自动切分的词素与人工切分的词素完全匹配的百分比。文[3]提出的方法词干提取准确度最高时达到 67%，比本文中的方法[2]提取准确度小于 30.66%。文[4]中的词干提取方法是对我们采用的方法[2]鲁棒性的提高。

表 4 给出了一些歧义的例子，这些例子只能通过句子等较长上下文中的形态学分析来消除其中的歧义。基于词内的词干方法无法可靠地提取词干。

表 4 有歧义的词干例子

变体	变体（拉丁文）	后缀
vAl（人民）/ val（拿）	vAl/val → vel+iN	iN
人的姓名/幸运的	qut → qutluq, qut+luq	luq
火/草	vot→ vot+ tAk, vot+laq	tAk, laq

通常，实验语料库中原始文本的字长可能不一样。因此，我们应该使用填充来修改文本长度，以使所有文本具有相同的字长，从而产生 LSTM 网络所需的矩阵。我们统计了我们语料库中每个原始文本中的单词数量，如图 5 所示。（在图 5 中，横轴表示文本中的单词数量，纵轴表示对应某个单词量的文本数量）。

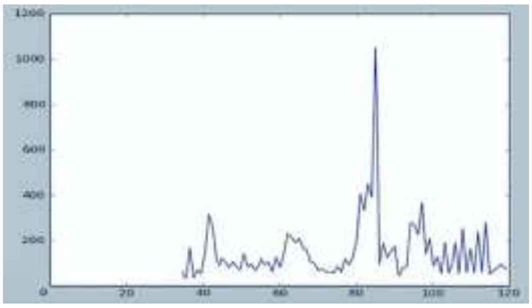

图 5 实验文本字长统计

从图 5 中可以看出，实验文本语料库中的文本词数趋向于约 40 字至 120 字范围之内，大部分文本约有 100 字左右。因此，我们选择了 100 个词作为用于 LSTM 的文本语料库的标准文本长度。我们用 0 填入字长不到 100 的文本。在这种情况下，我们将得到 LSTM 网络输入所需的文本矩阵。对于相应的词干序列文本，因为我们的切分器的词-词干标记比大致为 1/2，因此我们选择了 200 个词干单元作为 LSTM 网络的输入。

3 实验结果和分析

目前，维吾尔文本的分类研究还处于起步阶段，尚无公开可用的标准和开放的维吾尔语文本语料库可供我们进行特征提取和文本分类实验。因此，我们必须通过网络爬虫等技术来下载网上文本数据以构建维吾尔语文本语料库，以此进行实验。

3.1 实验语料库

我们使用网络爬虫技术从官方的维吾尔文网，如 uyghur.people.com.cn，下载文本来构建我们的文本语料库。我们的语料库包括法律、金融、体育、文化、卫生、旅游、教育、科学与娱乐等 9 大类，每类包含 500 篇，共 4500 篇。我们使用 75%的文本，是 3375 篇，作为训练文本语料，使用包括 450 篇的 10%的文本作为验证语料，其余部分作为测试语料。

针对网络文本容易出现拼写错误的情况，我们开发了维吾尔文字拼写检查工具。该工具通过分析维吾尔语音节的结构形式和规则，可以发现大部分有拼写错误的维吾尔

尔语词汇，从而我们能够更正给定维吾尔词汇中的拼写错误。拼写检查程序流程图如图6所示。

图 6 维吾尔文拼写检查程序流程

我们将所有文本从各种编码形式中规范化成统一的字符编码形式，并送入词素切分工具包中，转换成词干序列。基于词干的子词抽取方法能够很好地降低特征维数，其中词干词汇量的数量显著的下降到单词词汇量的31%以下，表5中所示。我们可以看出，随着类别数量和语料库数量的增加词干词汇的积累也只是词词汇积累的1/3。

表 5 词干提取引起的特征空间维数的减少

类别数	词词汇	词干词汇	词干-词词汇比率（%）
5	55165	18148	32.8
7	67924	21474	31.6
9	79762	24643	30.8

在稳健的语素切分之后，我们取每个文本的前200个词干，如果单词单位长度小于100，则用零填充较短的文本。然后使用CBOW算法获得所有语料库的词干向量。

3.2 评价指标

准确率、召回率和F1评分[25]用于评价文本分类。其中，准确率和召回率反映了分类的两个方面，F1是二者的结合。公式如下：

精确率=*准确被分类到类别* C_k *的文本数量/实际被分类到类别* C_k *的所有文本数量*

召回率=*准确被分类到类别* C_k *的文本数量/属于类别* C_k *的所有文本数量*

*F1=2*精确率*召回率/(精确率+召回率)*

对于我们所提出方法的评价，我们使用了宏观的F1测度。宏观的F1测度一个全局的F1指标。其中，首先分别计算每个类别的F1得分，然后将这些F1得分的算术平均值作为全局指标值。

3.3 实验结果与分析

我们将文本切分成词干序列，并选择200个词干单元，用word2vec对其词干进行向量化，主要采用基于KNN[5]、NB[7]、SVM[6]和LSTM的分类方法进行比较实验。前三种传统分类器上的分类实验中，用 x^2 统计方法，根据其词干项的CHI-2值大小选择了CHI-2值最大的前100到2000之间的若干特征维数以实现特征降维，进行了分类实验，实验结果如表6所示。

LSTM网络通过迭代计算获得权重，经多次迭代后得到理想的参数。在基于LSTM的分类试验中，我们形成100位的词（词干）向量作为LSTM网络的输入，为LSTM的层数选择了6个层，隐藏层的大小为64，用了MSE损失函数和Adam优化函数。我们将对基于100个词单元的分类结果与基于200个词干单元的分类结果进行了比较。我们做了150次迭代运算，如表7和8所示。

表 6 基于传统分类器的分类结果

CHI-2 特征词数	特征 Word2vec		
	分类器		
	KNN(%)	NB(%)	SVM(%)
100	77.12	82.67	84.78
200	79.36	84.29	87.10
400	81.70	86.37	88.59
600	82.27	88.24	91.30
800	84.03	91.62	92.05
1000	84.22	91.94	93.55
1500	84.47	91.32	92.90
2000	83.49	89.51	91.93

从表6、7和8可以看，基于KNN、NB和SVM的分类准确度最高时分别达84.47%、91.94%和93.55%。基于LSTM的实验中，在训练前段时间，随着迭代次数的增加，模型性能也随着增强，迭代次数达到40次左右时，模型基于词单元和词干单元的分类准确度都超过90%，并分别达到91.15%和93.57%。当迭代次数在60-70次左右时，

迭代对模型性能的影响开始下降，模型训练到饱和状态，模型基于词单元和词干单元的分类准确度达到 93.76%和 95.48%等峰值后开始收敛。与基于传统三种分类方法的分类准确度相比，本文提出的方法分类准确度分别高出 11.01%、3.54%和 1.93%。基于词干单元的分类准确度比基于词单元的分类准确度高出 2.72%。迭代次数增加时，训练时间也随着增加，但测试时间变化不大。

表 7 迭代次数对分类准确的的影响

实验性能	迭代次数									
	10	20	30	40	50	60	70	80	90	100
训练时间(s)	1021	2690	5622	7752	12513	17891	27610	32495	37451	42128
测试时间(s)	197	182	194	210	217	203	236	207	196	213
宏 F1(%)-词	83.44	85.93	88.27	91.15	92.47	92.94	93.18	93.11	93.22	93.48
宏 F1(%)-词干	86.73	89.96	92.62	93.57	94.89	95.19	95.23	95.33	**95.48**	95.11

表 8 迭代次数对分类准确的的影响

实验性能	迭代次数				
	110	120	130	140	150
训练时间(s)	47150	52640	57093	61724	67629
测试时间(s)	194	202	202	214	221
宏-F1(%)-词	**93.76**	93.51	93.54	93.42	93.43
宏-F1(%)-词干	95.23	95.25	95.36	95.28	95.19

为了验证本文提出的 word2vec 文本特征表示方法在文本分类任务中的性能，我们用 BOW 和 TFIDF 等传统的文本表示方法来表示文本特征（词特征和词干特征），将 LSTM 网络作为分类器，进行了文本分类实验，并对分类结果进行了比较，如表 9 所示。

表 9 基于不同文本表示方法的分类准确度

分类器	词汇单元	特征	宏 F1（%）	损失（%）
LSTM	词	BOW	92.21	13.11
		TFIDF	92.29	13.45
		Word2vec	93.76	11.89
	词干	BOW	93.91	13.19
		TFIDF	93.69	12.83
		Word2vec	95.48	11.42

从表 9 可以看出，基于 word2vec 的词单元和词干单元文本表示方法所得到的维吾尔文本分类准确度分别比基于传统文本表示方法的维吾尔文本分类准确度高出 1.55%， 1.47% 和 1.57%， 1.79%。基于 word2vec 的维吾尔文本分类损失值比基于传统的文本特征表示法的损失值小于 1.5%左右。由此可知，基于 word2vec 的深度文本表示方法能够有效的抓文本上下文之间的语义信息，能够更好的表示维吾尔文本及其特征，以提高文本的分类准确度。

4 结论

维吾尔语是一种形态丰富的粘着性语言，词是由缀加多个词缀所构成，这一性质在理论上造成无限的词汇量。后缀提供语义和句法功能。因此，词干提取和形态分析是自然语言处理的有效途径。谷歌开发的词向量技术可以将语言单元映射成基于上下文的顺序向量空间。从上下文信息中提取和预测 OOV 是一种自然的方法。本文讨论了一种基于词素并行数据的稳健词素切分方法，以及一种基于词（词干）向量和神经网络结构的文本分类方法。基于 LSTM 网络的模型，维吾尔文本分类任务分别在词和词干单元上实现。实验结果表明，与词单元相比，词干单元有多种优异的性质，更合适派生类语言的处理。

参考文献

[1] M. Ablimit, T. Kawahara, A. Hamdulla, et al. Stem-Affix based Uyghur Morphological Analyzer[J]. International Journal of Future Generation Communication and Networking, 2016, 9(2):59-72.

[2] Mijit Ablimit, Sardar Parhat, Askar Hamdulla, et al. Multilingual Language Processing Tool for Uyghur, Kazak and Kirghiz[C]// Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017:737-740.

[3] 早克热•卡德尔，艾山•吾买尔，艾斯卡尔•艾木都拉等. 维吾尔语名词构形词缀有限状态自动机的构造[J]. 中文信息学, 2009, 23(6):116-122.

[4] 赛迪亚古丽•艾尼瓦尔，向露，艾斯卡尔•艾木都拉等. 融合多策略的维吾尔语词干提取方法[J]. 中文信息学报, 2015, 29(5):204-211.

[5] Palidan Tuerxun, Fang Dingyi, Askar Hamdulla. The KNN based Uyghur Text Classification and its Performance Analysis[J]. International Journal of Hybrid Information Technology, 2015, 8(3):63-72.

[6] S. Imam, R. Parhat, A. Hamdulla, et al. Performance analysis of different keyword extraction algorithms for emotion recognition from Uyghur text[C]// International Symposium on Chinese Spoken Language Processing. IEEE, 2014.

[7] 陈洋，哈力旦•阿布都热依木，伊力亚尔•达吾提等. 基于加权改进贝叶斯算法的维吾尔文文本分类[J]. 计算机工程与设计，2014, 35(6):1999-2003.

[8] A. McCallum, K. Nigam. Text classification by bootstrapping with keywords, EM and shrinkage [C]// In ACL99-Workshop for Unsupervised Learning in Natural Language Processing, 1999:51-58.

[9] Y. Zhang, N. Zincir-Heywood, E. Millos. Narrative text classification for automatic key phrase extraction in web document corpora[C]// Proceedings of the 7th annual ACM international workshop on Web information and data management, 2005:52-58.

[10] 王汝娇，姬东鸿. 基于卷积神经网络与多特征融合的 Twitter 情感分类方法[J]. 计算机工程，2018, 44(2):210-219.

[11] B. Zhou, Y. Yao, J. Luo. Cost-sensitive three-way email spam filtering[J]. Journal of Intelligent Information Systems, 2014, 42(1) :19-45.

[12] C. C. Aggarwal, C. Zhai. A survey of text classification algorithms[J]. In Mining text data, 2012:163-222.

[13] XI. Xue-Feng, Z. Guo-Dong. A Survey on Deep Learning for Natural Language Processing[J]. Acta Automatica Sinica, 2016, 42(10):1445-1465.

[14] 赵淑芳，董小雨. 基于改进的 LSTM 深度神经网络语音识别研究[J]. 郑州大学学报（工学版）, 2018, 39(05):67-71.

[15] 张宇，张鹏远等. 基于注意力 LSTM 和多任务学习的远场语音识别[C]// 第十四届全国人机语音通讯学术会议（NCMMSC’2017）论文集. 2017.

[16] Wallach, M. Hanna. Topic modeling: beyond bag-of-words[C]// International Conference on Machine Learning. ACM, 2006:977-984.

[17] J. Hu, Y. Yao. Research on the Application of an Improved TFIDF Algorithm in Text Classification[J]. Journal of Convergence Information Technology, 2013, 8(7):639-646.

[18] Y. Bengio, H. Schwent, J. S. Senécal, et al. Neural Probabilistic Language Models[M]// Innovations in Machine Learning. Springer Berlin Heidelberg, 2006:211-219.

[19] A. Mnih, G. Hinton. Three New Graphical Models for Statistical Language Modelling[C]// Proceedings of 24th International Conference on Machine Learning. ACM, 2007:641-648.

[20] T. Mikolov, T. Sutskever, et al. Distributed Representation of Words and Phrases and Their Compo-sitionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.

[21] L. Siwei, Xu. Liheng, L. Kang, et al. Recurrent Convolutional Neural Networks for Text Classification[C]// Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015:2267-2273.

[22] Y. Goldberg, O. Levy. Word2vec Explained: deriving Mikolov et al.’s negative-sampling word-embedding method[J]. 2014, Eprint Arxiv.1402.3722.

[23] Y. Q. Chen, M. S. Nixon, R. I. Damper. Implementing the k-nearest neighbour rule via a neural network[C]// IEEE International Conference on

Neural Networks. IEEE, 1995:136-140.
[24] S. Hochreiter, J. Schmidhuber. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
[25] Sebastiani, Fabrizio. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002, 34(1):1-47.

作者联系方式：沙尔旦尔·帕尔哈提， 地址：新疆乌鲁木齐市胜利路 666 号（新疆大学信息科学与工程学院） 邮编：830046 电话：13999221222， 电子邮箱：sardar312@126.com