文章编号：1003-0077(2016)06-0182-08

限定领域口语对话系统中超出领域话语的对话行为识别

黄沛杰，王俊东，柯子烜，林丕源

（华南农业大学 数学与信息学院，广东 广州 510642）

摘 要：由于领域外话语具有内容短小、表达多样性、开放性及口语化等特点，限定领域口语对话系统中超出领域话语的对话行为识别是一个挑战。该文提出了一种结合外部无标签微博数据的随机森林对话行为识别方法。该文采用的微博数据无需根据应用领域特点专门收集和挑选，又与口语对话同样具有口语化和表达多样性的特点，其训练得到的词向量在超出领域话语出现超出词汇表字词时提供了有效的相似性扩展度量。随机森林模型具有较好的泛化能力，适合训练数据有限的分类任务。中文特定领域的口语对话语料库测试表明，该文提出的超出领域话语的对话行为识别方法取得了优于最大熵、卷积神经网络等短文本分类研究进展中的方法的效果。

关键词：对话行为识别；超出领域话语；随机森林；词向量；口语对话系统

中图分类号：TP391 **文献标识码**：A

Dialogue Act Recognition for Out-of-Domain Utterances in Spoken Dialogue System

HUANG Peijie, WANG Jundong, KE Zixuan, LIN Piyuan

(College of Mathematic and Informatics, South China Agricultural University, Guangzhou, Guangdong 510642, China)

Abstract: Due to the short length, diversity, openness and colloquial features of out-of-domain (OOD) utterances, such dialogue act (DA) recognition for OOD utterances remains a challenge in domain specific spoken dialogue system. This paper proposes an effective DA recognition method using the random forest and external information. The unlabeled Weibo dataset, which is not domain specific yet possesses the similar characteristic of colloquialism and diversity with the spoken dialogue, is used to train the word embedding by unsupervised learning method. The trained word embedding provides similar computing for out of vocabulary (OOV) words in the training and test OOD utterances. The evaluation on a Chinese dialogue corpus in restricted domain shows that the proposed method outperforms some state-of-the-art short text classification methods for DA recognition.

Key words: dialogue act recognition; out-of-domain utterance; random forest; word embedding; spoken dialogue system

1 引言

面向任务(task-oriented)的限定领域对话系统是目前人工智能和自然语言理解领域内的研究热点之一，已广泛应用于信息查询系统[1-5]、导航系统[6-7]、导游系统[8]和导购系统[9]等自然语言智能助理。然而，当使用自然语言对话时，即使用户了解某对话系统的限定领域，用户在对话流程中仍然不可避免会使用一些超出领域(out-of-domain，OOD)话语(utterance)，如问候、表态等[10]。事实上，OOD话语的现象很常见，例如，AT&T 的“How may I help you”系统[2]，以及 BTaxeCT 和 Lucent Bell 合作开发的“OASIS call-steering”系统[4]，大约有20%的用户问题是 OOD 的。尽管这些限定领域对话系统从完成任务角度上看只需要专注于特定的业务功能，但是如果能较为妥善地处理好 OOD 话语，而不仅仅是提示用户话语超出领域，将会有效地提高用户体验[11]。

对话行为(dialogue act，DA)识别是处理 OOD 话语的关键环节，是后续对话控制和应答的基础。在研究进展中，DA 识别通常被当作短文本分类问

收稿日期：2016-09-27 定稿日期：2016-10-20

基金项目：国家自然科学基金(71472068)；广东省大学生科技创新培育专项项目(pdjh2016b0087)

题[12]。然而，与评论等短文本信息相比，限定领域口语对话系统中的 OOD 话语通常长度更短，也更为口语化，并且比领域内(in-domain)话语更具开放性和表达多样性，其 DA 的有效识别仍然是个挑战。已有的 OOD 话语相关研究工作主要集中在 OOD 话语的检测，并根据检测结果简单响应用户，而缺少对 OOD 话语 DA 的有效识别[7,13-15]。

本文提出一种结合外部无标签数据的 OOD 话语 DA 识别方法。由于训练分类模型的 OOD 话语样例数量有限，以及 OOD 话语的语义开放性和口语表达多样性，待分类 OOD 话语中有时会出现超出词汇表(out-of-vocabulary，OOV)字词。而同样具有口语化和表达多样性的微博数据的“字词相似性”可预期能接近于限定领域口语对话系统中 OOD 话语的“字词相似性”。因此，本文采用分布式表达方式训练无标签微博数据得到词向量(word embedding)，并用于帮助待分类 OOD 话语出现 OOV 字词时实现有效的特征扩充。分类模型采用了随机森林(random forests，RF)模型[16]，并通过交叉验证的方式进行了参数选择。相比于已有的研究，本文的主要贡献包括：

(1) 采用无标签微博数据训练的词向量作为相似性度量，在待分类 OOD 话语出现 OOV 时提供 OOV 字词的相似性扩展，从某种程度上解决了 OOD 话语的开放性带来的对话语料库词汇覆盖不全的问题，也增强了识别方法对 OOD 所固有的口语化和表达多样性的适应。

(2) 在中文手机导购领域的对话系统中评测了基于随机森林的 OOD 话语 DA 识别方法，在训练数据有限的情况下，取得了优于最大熵(maximum entropy，ME)、卷积神经网络(convolutional neural network，CNN)等短文本分类研究进展中的方法的效果。

本文后续部分安排如下：第二节介绍相关工作；第三节介绍本文提出的方法；第四节给出测试结果及分析；最后，第五节总结本文的工作并做简要的展望。

2　相关工作

在短文本信息，如微博、商品评论、影评等的分析领域，为了克服短文本具有的噪音多、特征稀疏和主题不明确等特点[17]，许多机器学习模型如 SVM (support vector machine)[18]、最大熵[19]、CNN[20]被应用于短文本分类。此外，为了解决短文本分类问题中数据稀疏问题，结构化语义知识库如 Wikipedia、WordNet 等常被用于语义相似性计算[21]，另外一些研究则采用在领域相关的无标签数据集上使用 LDA(latent dirichlet allocation)获取主题特征[22]或者使用神经网络(neural network)训练词向量[19]的方法增加语义特征。

在口语对话系统领域内话语的 DA 识别方面，传统的语言模型和机器学习方法如 N-gram[23]、朴素贝叶斯(naïve bayes)[24]、决策树(decision tree)[25]、最大熵[26]、神经网络[27]、隐马尔科夫(hidden markov model，HMM)[28]、条件随机场(conditional random field，CRF)[29]等各种分类模型被应用。较为丰富的语义或语法等文本信息被良好表达并输入到分类模型。有些研究还考虑了对话上下文的序列信息[28-29]以及更深层次的异构特征学习[29]。

上述研究进展的方法都对 OOD 话语的 DA 识别提供了很好的借鉴。然而与短文本分类及领域内话语的 DA 识别相比，OOD 话语的 DA 识别具有以下挑战：

(1) 口语对话系统话语比微博、评论、新闻标题等常见短文本信息更短。如搜狗实验室提供的中文新闻标题分类数据集，大部分文本数据长度集中在 10～21 字之间[19]，微博、电影评论等的平均长度则更长一些，而在我们实验中的对话语料，OOD 话语平均长度只有 3.6 字，集中在 1～8 个字之间。短文本所固有的噪音多和特征稀疏在口语对话的 OOD 话语中表现得更为突出。另外，口语对话中的 OOD 话语比电影评论和新闻标题等短文本更为口语化，比微博也多了一些口语化省略的情况。

(2) 相比于领域内话语，OOD 话语语义更为开放和表达多样，容易产生 OOV 字词，并且也缺少领域内话语携带的相对较为丰富的语义或语法等文本信息。此外，OOD 话语与对话上下文的关联也远远没有领域内话语高。

王俊东等人[10]提出的 OOD 话语处理方案中也包含了对 OOD 话语的 DA 识别，采用了向量空间模型(vector space model，VSM)，通过词频和期望交叉熵(expected cross entropy，ECE)权重计算句子相似度，不足之处在于 DA 识别方法比较简单，并且缺乏对 OOD 话语中 OOV 字词的考虑。本文采用大量无标签微博数据训练的词向量作为相似性度量，为 OOD 话语中的 OOV 字词提供相似性扩展。与大多数文本分类研究相比，本文并没有依赖于领

域密切相关的外部数据，更易于实现。此外，在中文训练语料数据有限的情况下，考虑到以决策树为基学习器构建 Bagging 集成的随机森林模型在小样本数据集上的良好表现[16]，本文采用随机森林作为分类模型。

3 基于随机森林和外部数据词向量的对话行为识别方法

3.1 总体技术架构

图1是本文提出的方法的总体技术架构。

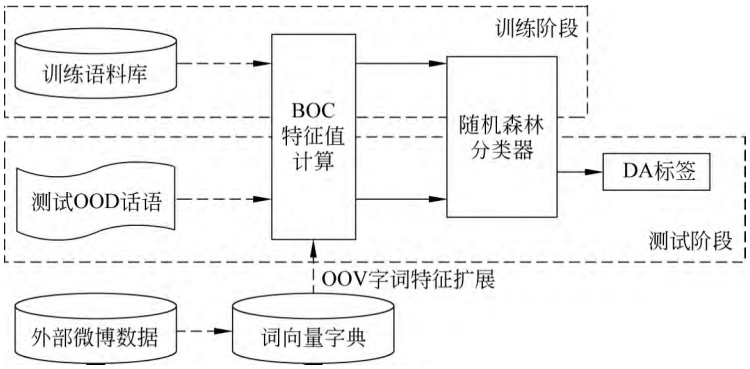

图1 限定领域口语对话系统 OOD 话语 DA 识别方法技术架构

在这个架构中，主要分成两个阶段：(1)在训练阶段，将训练语料库中句子进行预处理，进而针对对话系统 OOD 话语口语化的特点，利用 BOC(bag of Chinese characters)计算特征值，生成特征向量，并使用随机森林分类器进行模型训练；(2)在测试阶段，首先对检测到出现 OOV 字词的待识别 OOD 话语进行相似性扩展。相似性扩展通过计算字词间词向量的余弦相似度，找出 OOV 字词最相近的训练语料中的字词扩展 OOD 话语。接着将扩展后的 OOD 话语进行 BOC 特征值计算，并生成特征向量。最后使用由(1)训练得到的分类器进行 DA 标签的分类。

3.2 外部数据词向量

词向量通常被称为“word representation”或“word embedding”，是通过训练无标签语料将每个词映射成低维实数向量的方法，每一维都代表了词的浅层语义特征[30]，通过低维实数向量之间的距离(例如余弦相似度、欧式距离等)来描述字词之间的语义相似度。低维的词向量避免了用传统的稀疏表达在解决某些任务的时候(比如构建语言模型)所造成的维数灾难[31]。本文采用与 OOD 话语同样具有口语化和表达多样性的微博数据来训练词向量。

目前训练词向量的主流方法是在训练语言模型的同时得到词向量。基于统计的语言模型能够表示成一个已出现的词和当前词的条件概率的极大似然估计为式(1)。

$$\hat{P}(w_1^T)=\prod_{t=1}^{T}\hat{P}(w_t\mid w_1^{t-1}) \tag{1}$$

其中，w_t 表示句子中第 t 个词，$w_i^j=(w_i,w_{i+1},\ldots,w_j)$ 表示句子中下标 i 到 j 的子串。

针对不同的上下文构造方法，在训练词向量时主要有 CBOW (continuous bag-of-words)和 Skip-gram 两种语言模型[32]。Skip-gram 模型允许某些词被跳过，在训练数据少的情况用 Skip-gram 可以创造更多的训练例子，而连续的 CBOW 则可以有较快的训练速度[32]。由于本文采用的是大量微博数据，因此本文使用 CBOW 语言模型对词语的语义层面建模。CBOW 语言模型不仅限于已出现的词为 w_t 的上下文，而是考虑了句子中距离当前词为 n 以内的词都看作是当前词的上下文环境，如图2所示。

用一个函数 f 表示当前词 w_t 的上下文的向量到当前词 w_t 条件概率的映射[31]，并结合 CBOW 的机制，则当前词的上下文和当前词的条件概率可以表示为式(2)。

$$\begin{aligned}&\hat{P}(w_t\mid context(w_t))\\&=f(w_t,C(w_{t-n}),\ldots,C(w_{t-1}),C(w_{t+1}),\ldots,C(w_{t+n}))\\&=f(w_t,\sum_{0<|i-t|\leqslant n}C(w_i))\end{aligned} \tag{2}$$

其中，$C(w_i)$ 是词语 w_i 的分布式特征向量。

在训练语言模型及词向量时，对于 w_t 都要扫一遍词库大小 $|V|$，计算复杂度过高。可以采用负采样(negative sampling)[33]和分层 softmax(hierar-

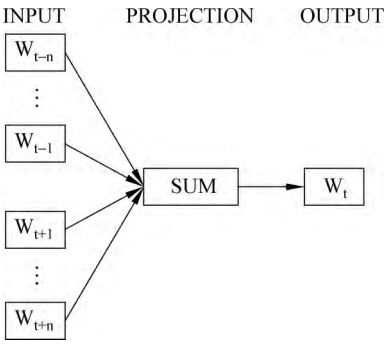

图 2　CBOW 语言模型架构

chical softmax)[34]的方法来降低计算复杂度。

3.3　随机森林模型

随机森林(random forest)[16]作为一种集成学习(ensemble learning)方法，是一种利用多棵树为基学习器构建 Bagging 集成的分类器。本文采用了 Breiman 提出的基于分类回归树(classification and regression trees，CART)[35]的随机森林模型。该模型具有良好的实用性能和处理高维数据的能力，并且只依赖于少数的几个容易调节的参数，已成为模式识别问题的一种常用的学习算法[36]。随机森林模型结合了 Breiman 的自助聚集(bootstrap aggregating)[37]思想和 Ho 的随机子空间(random subspace)[38]方法，其模型训练原理[39]如图 3 所示。其中，k 对应随机森林的子树数量，子树的分裂次数N由不同子树的样本和特征决定。每颗子树都分裂直至最大生长，即同一个节点下所有训练样例都属于同一个类别。

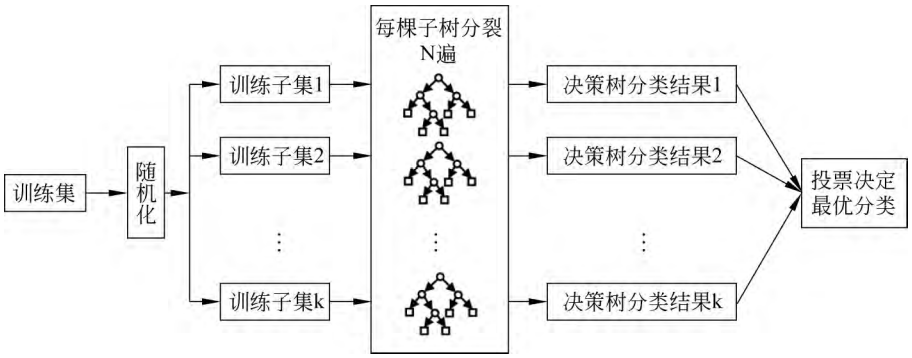

图 3　随机森林训练原理示意图

大量的理论和实证研究都证明了随机森林模型具有很高的预测准确率，对异常值和噪声具有很好的容忍度，且不容易出现过拟合[39]。

4　实验

4.1　数据集

训练数据集采用了和文献[10]一样的 OOD 话语训练库，共 1 238 句。我们在文献[10]的基础上进一步完善了 DA 分类，如表 1 所示。共五大类(维度)25 小类(交互功能)。

值得注意的是，尽管在一些文献中，OOD 话语只限于身份信息、天气等闲聊话语，在我们的研究中，我们把不携带领域语义信息的用户话语都当成 OOD 话语。这个广义的 OOD 定义使得一些领域任务相关的话语也被归类到 OOD 话语，如肯定或者否定的表态，或者惯用开场语。我们希望这样的 OOD 定义对限定领域口语对话系统是有益的，因为携带领域语义信息的话语可以根据领域语义得到较

表 1　OOD 话语的 DA 类别与示例

对话行为		OOD 话语示例
维度	交互功能	
任务进程	开始	我想买手机。
	结束	再见。
	更换	换一款。
	成交	买了。
	不成交	我不想买了。
	详情	具体点。
	不理解	啥意思？
表态	肯定	好的。
	否定	不是。
	疑问	真的吗？
	满意	挺好的！
	不满	差评！
	附和	呃...
	犹豫	不知道哪个好。
	随便	无所谓。

续表

对话行为		OOD话语示例
维度	交互功能	
社交义务	问候	你好！
	致谢	谢谢！
	道歉	不好意思。
	接受致谢	不用谢。
	接受道歉	没关系。
闲聊	时间	现在几点啦？
	天气	今天天气好冷。
	身份信息	你叫什么名字？
其他	骂人	混蛋！
	其他	你猜。

好的处理。此外，我们用一个“其他”小类代表不属于任何前24个小类的OOD话语，该小类的训练集只用于匹配，而不参与识别模型的建模。

我们在实现的中文手机导购对话系统[9]中进行了测试。系统的测试人员是15名学生志愿者，每位测试者测试12～14段。由于本文关注的是OOD话语的DA分类，因此，没有正常结束的对话（可能是系统异常中断或者用户异常退出连接）中的OOD话语也可以使用。

用于测试的对话语料的总体情况如表2所示。

表2 测试语料的情况

总对话段数	用户话语总数	OOD话语数量（去重前/去重后）	待识别的OOD话语数量
193	2 070	362/166	131

对话语料库共193段对话，用户话语总数为2 070，OOD的数量为362，占了17.5%，与文献[2]和[4]中的口语对话系统的OOD比例相似，表明了OOD识别在限定领域口语对话系统研究和应用中的价值。在131例未被训练集覆盖的待识别的OOD话语中，有四例属于其他小类。因此，本文的测试集即为去除了四例其他小类之后的127句OOD话语。

4.2 实验设置

本文的外部数据库采用的是中国中文信息学会社会媒体专委会提供的SMP2015微博数据集（SMP 2015 Weibo DataSet）。该数据集超过500G，目前我们采用了其中的一个子集（1 000万条微博，519 734词汇，约1.5G），与相关方法采用的搜狗实验室新闻数据（Sougo News）（515 789词汇）具有相当的词汇量标准。我们也验证过更大的微博数据量，在当前的DA识别任务中没有显著的识别效果提升。词向量采用Python Gensim主题模型包中的word2vec进行训练。随机森林和CNN模型的参数通过K-折（本文的实验采用3折）交叉验证得到。

实验方案为：

（1）随机森林模型的参数选择：验证不同的子树数量的随机森林模型的性能；

（2）原始特征的选择：对比字和词作为原始特征的DA识别效果；

（3）研究进展方法DA识别性能对比：对比了本文提出的方法与研究进展方法的DA识别结果。并对比了不同外部数据对OOV相似性扩展的效果；

（4）训练库规模的影响：采用不同比例的训练语料库，验证本文提出的方法对训练数据规模的依赖性。

本文的方法，结合外部无标签微博数据训练的词向量为度量的OOV相似性扩展的随机森林模型，记为RF(BOC+OOV(w2v))，对比的三种研究进展的方法如下。

（1）VSM(ECE)：王俊东等人[10]应用于OOD话语DA分类的方法，利用ECE选出类别特征词，并将类别特征词以VSM向量形式表示类别，通过词频和ECE权重计算句子相似度；

（2）ME(TFIDF)：马成龙等人[19]应用于短文本（网页搜索片段和新闻标题）分类的方法，对训练数据所生成的词典利用TFIDF计算特征值，采用最大熵模型进行分类；

（3）CNN(w2v)：Kim[20]应用于短文本（电影评论等）分类的方法，采用Google新闻语料训练得到的词向量表达短文本中的词语，分类模型采用了CNN，并使用了3、4、5三种不同卷积窗口的卷积核。在本文的实验中，我们采用微博数据训练得到的词向量训练CNN模型，并通过交叉验证选择最优的卷积核数量。

4.3 实验结果分析

4.3.1 随机森林模型的参数选择

随机森林的关键参数包括子树的棵数k以及每棵树随机选取的特征数m。通过交叉验证发现m

的最优值与经验公式 log (*M*) 相近，其中 *M* 为总特征数，本文实验中 *M* 为 754。不同的子树数量的随机森林模型的训练和验证结果如图 4 所示，采用的识别方法是本文的 RF(BOC+OOV(w2v))方法。

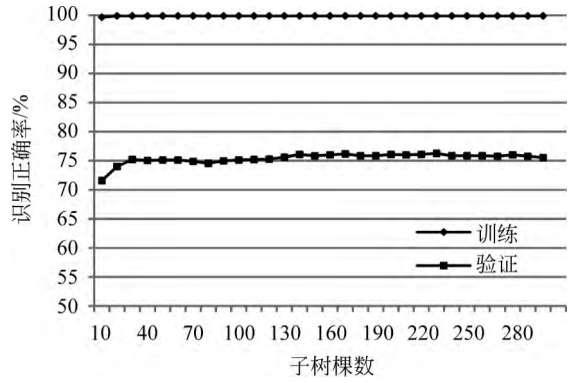

图 4　不同子树数量的随机森林训练和验证结果

可以看到随机森林模型在训练误差已经接近为 0(20 棵子树)的情况下，随着子树数量进一步增加，模型并没有马上进入过拟合状态，其交叉验证的正确率继续保持提升。另一方面，也可以看到，不需要太复杂的模型(140 棵子树左右)就可以接近性能上限(在现有的数据集条件下)，并且随着子树的进一步增加保持了较稳定的验证结果，不容易产生模型过拟合。

4.3.2　原始特征的选择

我们对比了各种模型选用字和词作为原始特征的 DA 识别效果，如图 5 所示。

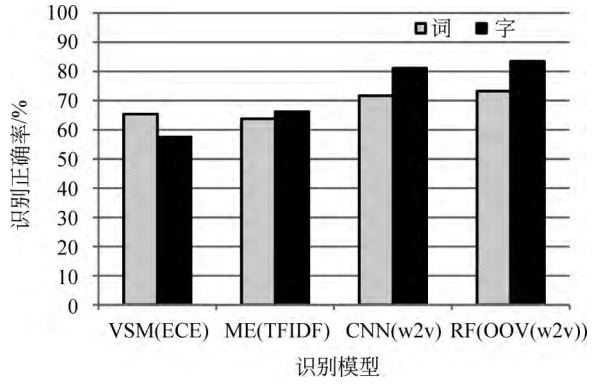

图 5　不同原始特征(词和字)的 DA 识别效果

从图 5 可以看到，除了 VSM 模型，其他模型采用字为原始特征的识别效果比采用词的好，尤其是 CNN 和 RF 模型，这也反映了对话系统的 OOD 话语口语化的特点。VSM 模型词比字作为原始特征的识别效果更好，可能是因为模型简单，未能很好地实现由字到词的特征搭配。

4.3.3　研究进展方法 DA 识别性能对比

本文的方法与研究进展方法的 DA 识别结果如表 3 所示。根据图 5 的对比，除了 VSM 模型采用词为原始特征，其他方法都采用了字为原始特征。其中，我们也对比了使用不同外部数据训练的词向量作为 OOV 字词提供相似性扩展时的度量的效果。为了区别采用搜狗实验室的新闻数据(Sougo News)训练的词向量作为距离度量的方法，在表 3 中，本文的方法标记为 RF(BOC+OOV(w2v_SMP-Weibo))。在本文的其他比较中，本文的方法标记为 RF(BOC+OOV(w2v))。

表 3　本文方法与研究进展方法的 DA 识别对比

方法	识别正确率/%
VSM(ECE)	65.35
ME(TFIDF)	66.14
CNN(w2v)	81.10
RF(BOC)	81.89
RF(BOC+OOV(w2v_Sougo-News))	81.89
RF(BOC+OOV(w2v_SMP-Weibo))	84.25

从表 3 可以看到，本文提出的方法比 VSM(ECE)、ME(TFIDF)和 CNN(w2v)等方法分别提高了 18.90%、18.11%和 3.15%的 OOD 话语 DA 识别正确率。与口语对话系统 OOD 话语同样具有口语化和表达多样性的微博数据(SMP-Weibo)训练的词向量作为距离度量能更好地为 OOD 话语中的 OOV 字词提供合适的相似性扩展，而采用搜狗实验室的新闻数据(Sougo News)训练的词向量作为距离度量没能帮助提高识别正确率。我们还进一步对比了 RF 和 CNN 方法的识别稳定性，采用了在模型选择时的验证 Top 5 的模型在测试集上的 DA 识别正确率进行对比，如图 6 所示。

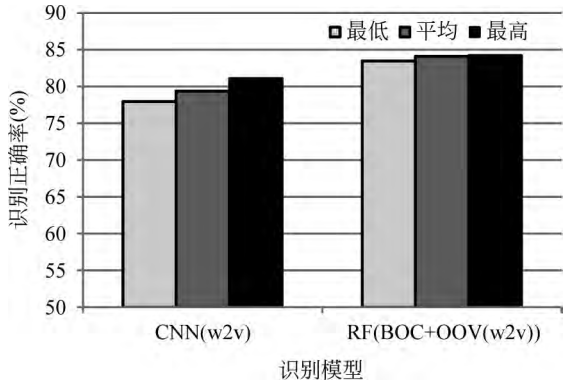

图 6　不同识别模型的识别性能稳定性

从图6可以看到，我们的方法比CNN模型具有更稳定的识别效果。我们的方法的验证Top 5的模型对应的最低、平均和最高测试正确率分别为83.46%、84.09%和84.25%，而CNN的验证Top 5的模型对应的最高和最低的测试正确率差异则超过3%。

4.3.4 训练库规模的影响

我们进一步验证了本文提出的方法对训练数据规模的依赖性。我们保持DA类别分布比例不变，将训练语料库平均分成十份，每次随机增加一份作为训练数据。共进行了十遍实验(选择不同的一份作为第一份)。使用同样的测试集进行测验，测验的方法包括本文的RF(BOC+OOV(w2v))方法以及没对OOV进行相似性扩展的RF(BOC)方法，结果如图7所示。从图中结果可以看到，随着训练语料库规模的增大，两种方法的识别正确率都保持增长，可见DA识别方法对训练语料的依赖还是比较大的。另一个方面也可以看到，目前规模的训练语料的50%已经可以使本文的识别方法获得较好的识别正确率(70%+)。

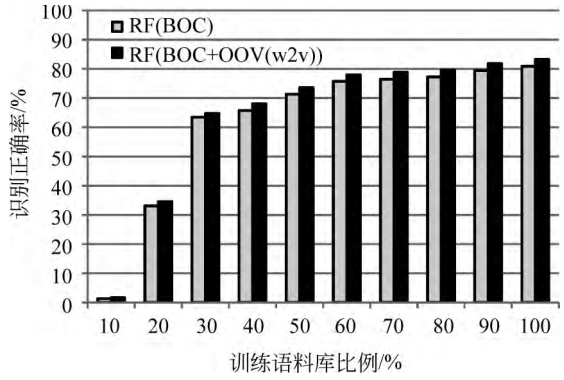

图7 不同比例训练语料的测试结果

5 结束语

本文基于外部无标签微博数据训练的词向量和随机森林模型，提出了一种限定领域口语对话系统OOD话语的DA识别方法。在中文手机导购领域的OOD话语测试表明，本文的方法取得了优于研究进展中的短文本分类方法的应用效果。与限定领域口语对话系统中OOD话语同样具有口语化和表达多样性特点的微博数据训练得到的词向量，有助于为待分类的OOD话语中的OOV字词找到合适的近似扩展。随机森林模型在有限的OOD话语训练数据集的条件下，取得了优于最大熵和CNN等模型的识别效果。未来计划通过分析存在的识别错误样例，并通过人工标注对话语料中的OOD话语，结合进一步扩大的训练库，探索CNN和长短期记忆人工神经网络(long-short term memory，LSTM)等具有一定结构化学习优势的模型在OOD话语的DA识别中性能提升的可能，以及多种识别模型有效结合的方法。

参考文献

[1] Price P J. Evaluation of spoken language systems: the ATIS domain[C]//Proceedings of DARPA Workshop on Speech and Natural Language, Hidden Valley, PA, 1990.

[2] Gorin A, Riccardi G, Wright J. How may I help you? [J]. Speech Communication, 1997, 23(1-2): 113-127.

[3] Zue V, Seneff S, Glass J, et al. JUPITER: a telephone-based conversational interface for weather information[J]. IEEE Transactions on Speech and Audio Processing, 2000, 8(1): 85-96.

[4] Durston P, Farrell M, Attwater D, et al. OASIS natural language call steering trial[C]//Proceedings of 7th European Conference on Speech Communication and Technology (Eurospeech 2011), 2001: 1323-1326.

[5] 张琳，高峰，郭荣，等. 汉语股票实时行情查询对话系统[J]. 计算机应用，2004，24(7)：61-63.

[6] 黄寅飞，郑方，燕鹏举，等. 校园导航系统EasyNav的设计与实现[J]. 中文信息学报，2001，15(4)：35-40.

[7] Reichel C S, Sohn J, Ehrlich U, et al. Out-of-domain spoken dialogs in the car: a WoZ study[C]//Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2014), 2014: 12-21.

[8] Pappu A, Rudnicky A. The structure and generality of spoken route instructions[C]//Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2012), 2012: 99-107.

[9] Huang P J, Lin X M, Lian Z Q, et al. Ch2R: a Chinese chatter robot for online shopping guide[C]//Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2014), 2014: 26-34.

[10] 王俊东，黄沛杰，林仙茂等. 限定领域口语对话系统中超出领域话语的协处理方法[J]. 中文信息学报，2015，29(5)：194-203.

[11] Ameixa D, Coheur L, Fialho P, et al. Luke, I am your father: dealing with out-of-domain requests by

using movies subtitles [J]. IVA 2014. LNCS (LNAI), vol. 8637, pp. 13-21. Springer, Heidelberg (2014)

[12] Novielli N. and Strapparava C. The role of affect analysis in dialogue act identification [J]. IEEE Transactions on Affective Computing, 2013, 6(1): 1-14.

[13] Lane I R, Kawahara T, Matsui T, et al. Out-of-domain utterance detection using classification confidences of multiple topics[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(1): 150-161.

[14] Tür G, Deoras A, Hakkani-Tür D. Detecting out-of-domain utterances addressed to a virtual personal assistant[C]//Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014), 2014: 283-287.

[15] Celikyitmaz A, Hakkani-Tür D, Tür G. Approximate inference for domain detection in spoken language understanding [C]//Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), 2011: 1293-1296.

[16] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1), 5-32.

[17] Chen M G, Jin X M, Shen D. Short text classification improved by learning multigranularity topics [C]//Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011), 2011: 1776-1781.

[18] Silva J, Coheur L, Mendes A C, et al. From symbolic to sub-symbolic information in question classification. Artificial Intelligence Review, 2011, 35(2): 137-154.

[19] **马成龙，姜亚松，李艳玲，等. 基于词矢量相似度的短文本分类[J]. 山东大学学报：理学版**, 2014(12): 18-22.

[20] Kim Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 2014: 1746-1751.

[21] Kenter T, Rijke M D. Short text similarity with word embeddings[C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM 2015), 2015: 1411-1420.

[22] Phan X H, Nguyen L M, Horiguchi S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections[C]//Proceedings of the 17th International World Wide Web Conference (WWW 2008), 2008: 91-100.

[23] Louwerse M M, Crossley S A. Dialog act classification using n-gram algorithms[C]//Proceedings of 19th Florida Artificial Intelligence Research Society Conference (FLAIRS 2006), 2006: 758-763.

[24] Levin L, Langley C, Lavie A, et al. Domain specific speech acts for spoken language translation[C]//Proceedings of 4th SIGdial Workshop on Discourse and Dialogue (SIGDIAL 2003), 2003.

[25] Irie Y, Matsubara S, Kawaguchi N, et al. Speech intention understanding based on decision tree learning [C]//Proceedings of 8th International Conference on Spoken Language Processing (INTERSPEECH 2004-ICSLP), 2004.

[26] Lan K C, Shiu H K, Pong Luk Robert Wing, et al. Dialogue act recognition using maximum entropy[J]. Journal of the American Society for Information Science & Technology, 2008, 59(6): 859-874.

[27] Král P, Cerisara C, Klecková J. Combination of classifiers for automatic recognition of dialog acts[C]//Proceedings of 9th European Conference on Speech Communication and Technology (INTERSPEECH 2005-Eurospeech), 2005: 825-828.

[28] Lee S, Seo J. Korean speech act analysis system using hidden markov model with decision trees[J]. International Journal of Computer Processing of Oriental Languages, 2002, 15(03): 231-243.

[29] Zhou Y, Hu Q, Liu J, et al. Combining heterogeneous deep neural networks with conditional random fields for chinese dialogue act recognition[J]. Neurocomputing, 2015, 168(C): 408-417.

[30] Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), 2010: 384-394.

[31] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, (3): 1137-1155.

[32] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in Vector Space[C]//Proceedings of the 1st International Conference on Learning Representations (ICLR 2013), 2013.

[33] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [C]//Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS 2013), 2013: 3111-3119.

[34] Morin F, Bengio Y. Hierarchical probabilistic neural network language model[C]//Proceedings of the International Workshop on Artificial Intelligence and Statistics (AISTATS 2005), 2005: 246-252.

（**下转第** 200 **页**）

阿依古丽·哈力克(1991—),硕士,主要研究领域为自然语言处理与机器翻译。
E-mail: 1506867752@qq.com

艾山·吾买尔(1981—),通信作者,副教授,主要研究领域为自然语言处理与机器翻译。
E-mail: hasan1479@xju.edu.cn

吐尔根·伊布拉音 (1958—),教授,主要研究领域为自然语言处理、机器翻译、软件工程。
E-mail: turgun@xju.edu.cn

(上接第 189 页)

[35] Breiman L, Friedman J, Olshen R A, et al. Classification and regression trees[M]. Chapman & Hall, New York, 1984.

[36] Scornet E. Random forests and kernel methods [J]. IEEE Transactions on Information Theory, 2015, 62(3): 1485-1500.

[37] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 26(2): 123-140

[38] Ho, T K. The random subspace method for constructing decision forests[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1998, 20(8), 832-844.

[39] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-38.

黄沛杰(1980—),通信作者,博士,副教授,主要研究领域为人工智能、自然语言处理、口语对话系统。
E-mail: pjhuang@scau.edu.cn

王俊东(1992—),硕士研究生,主要研究领域为自然语言处理。
E-mail: jdwang@stu.scau.edu.cn

柯子烜(1995—),本科生,主要研究领域为自然语言处理。
E-mail: iscauzixuanke@gmail.com