智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

面向鲁棒口语理解的声学组块混淆语言模型微调算法

李荣军, 郭秀焱, 杨静远

引用本文:
李荣军,郭秀焱,杨静远. 面向鲁棒口语理解的声学组块混淆语言模型微调算法[J]. 智能系统学报, 2023, 18(1): 131–137.
LI Rongjun,GUO Xiuyan,YANG Jingyuan. A fine-tuning algorithm for acoustic text chunk confusion language model orienting to understand robust spoken language[J]. *CAAI Transactions on Intelligent Systems*, 2023, 18(1): 131–137.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202109024

DOI: 10.11992/tis.202109024
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.tp.20220930.1621.006.html

面向鲁棒口语理解的声学组块混淆语言模型微调算法

李荣军, 郭秀焱, 杨静远
(华为技术有限公司 AI 应用研究中心, 广东 深圳 518129)

摘 要: 利用预训练语言模型(pre-trained language models, PLM)提取句子的特征表示, 在处理下游书面文本的自然语言理解的任务中已经取得了显著的效果。但是, 当将其应用于口语语言理解(spoken language understanding, SLU)任务时, 由于前端语音识别(automatic speech recognition, ASR)的错误, 会导致 SLU 精度的下降。因此, 本文研究如何增强 PLM 提高 SLU 模型对 ASR 错误的鲁棒性。具体来讲, 通过比较 ASR 识别结果和人工转录结果之间的差异, 识别出连读和删除的文本组块, 通过设置新的预训练任务微调 PLM, 使发音相近的文本组块产生类似的特征嵌入表示, 以达到减轻 ASR 错误对 PLM 影响的目的。通过在 3 个基准数据集上的实验表明, 所提出的方法相比之前的方法, 精度有较大提升, 验证方法的有效性。

关键词: 自然语言理解; 口语语言理解; 意图识别; 预训练语言模型; 语音识别; 鲁棒性; 语言模型微调; 深度学习
中图分类号: TP18 **文献标志码**: A **文章编号**: 1673−4785(2023)01−0131−07

中文引用格式: 李荣军, 郭秀焱, 杨静远. 面向鲁棒口语理解的声学组块混淆语言模型微调算法 [J]. 智能系统学报, 2023, 18(1): 131–137.
英文引用格式: LI Rongjun, GUO Xiuyan, YANG Jingyuan. A fine-tuning algorithm for acoustic text chunk confusion language model orienting to understand robust spoken language[J]. CAAI transactions on intelligent systems, 2023, 18(1): 131–137.

A fine-tuning algorithm for acoustic text chunk confusion language model orienting to understand robust spoken language

LI Rongjun, GUO Xiuyan, YANG Jingyuan
(AI Application Research Center, Huawei Technologies Co., Ltd., Shenzhen 518129, China)

Abstract: Employing the pre-trained language model (PLM) to extract the feature representations of sentences has achieved remarkable results in processing downstream natural language understanding tasks based on texts. However, when applying PLM to spoken language understanding (SLU) tasks, it shows degraded performance resulting from erroneous text from front-end automatic speech recognition (ASR). To address this issue, this paper investigates how to enhance a PLM for better SLU robustness against ASR errors. Specifically, by comparing the differences between ASR recognition and manual transcription results, we identify the concatenated and deleted text chunks. Then, we set up a new pre-training task to fine-tune the PLM to make text chunks with similar pronunciation produce similar feature embedding representations to reduce the influence of ASR errors on PLM. Experiments conducted on three SLU benchmark datasets validate the efficiency of our proposal by showing significant accuracy improvements through comparison with prior arts.

Keywords: natural language understanding; spoken language understanding; intent recognition; pre-trained language model; speech recognition; robust; fine-tuning of language model; deep learning

口语语言理解作为任务型对话系统的核心组件, 目的是从语音识别文本中获取用户的意图表示信息, 并将这些信息提供给对话管理组件进行分析和决策。近年来, 随着深度学习技术的引入, ASR 和 SLU 技术得到了长足发展, 精度获得显著提升[1-3]。尽管如此, 当有域外语音或噪声语音输入时, ASR 文本中会存在插入、删除和修改错误, 降低意图识别的准确率, 影响对话系统的用户体验[4-5]。比如句子“Add song too sleepy dime”的正确用户意图应该是“Add to playlist”, 但这个句子被 ASR 误识为“I 'd song too sleepy dime”时, 它的意图变成了“Search screening event”。在这个例子中, 因为“Add”被误识别成“I’d”, 用户意图就发生了漂移。

为了应对这类挑战, 有文献提出从语音信号到意图的端到端模型, 直接从语音中识别意图[6-9]。文献 [10] 将成熟 ASR 模型的声学模型部分作为预训练的语音编码器, 然后联合微调该语音编码器和意图识别任务, 提升了识别精度。文献 [11] 提出学习音素加文本的预训练语言模型, 基于它

收稿日期: 2021−09−13. **网络出版日期**: 2022−10−18.
通信作者: 李荣军. E-mail: lirongjun3@huawei.com.

的意图识别模型对转录错误的容忍度有一定提升。但上述方法都需要除文本外的资源支持，即音素和语音信号，限制了使用场景。

另外有文献提出从发音混淆词入手，提高意图识别模型的鲁棒性。此方向的工作分为：1）对ASR 解码前的格结构（lattice）进行建模的方法[12-15]。格是一种有向图结构，包含多条解码路径，每条路径对应一种可能的文本转录结果。解码路径间存在替代词、发音混淆词序列。文献 [14] 中提出 LatticeRNN 模型，沿着格的边结构编码和汇总所有节点信息，使得终止节点的编码结果可代表格中多条转录结果。由于更充分地对识别结果进行表示，能够提高意图模型的精度，文献 [15] 对上述方法进行扩展，首先训练格语言模型（lattice language modeling, LLM），再使用任务数据微调语言模型得到更好的意图识别模型。2）对 ASR 多候选（N-best）进行建模的方法。ASR 在解码过程中，通过控制集束搜索的参数，可以获得多候选的识别结果。文献 [16] 使用该结果作为模型输入，联合训练语言纠错任务和意图识别任务，相比对单候选输入进行建模的方法，该类方法都能通过更充分的信息进行决策。3）对混淆样本对进行建模的方法[17-19]。它们通过改善预训练语言模型对发音混淆样本的鲁棒性，提高意图识别精度。文献 [18] 通过对比误识文本和正确文本的差异，寻找发音混淆词汇对。指导语言模型微调，促使模型对这些词汇对产生相似的编码。文献 [19] 使用随机插入和删除操作来制造混淆词汇对，并训练扭曲语言模型，提高它对插入和删除词汇的预测能力。由于这类方法不需要读取 ASR 内部状态，词汇对的构造过程简单、构造代价低，更适合用在工业场景。但是，现有工作只利用了混淆词汇对，而没有考虑普遍的连读误识的情况，即单个词被误识别多个词，比如 goddess 被识别成 got 和 us，反之亦然。本文扩展了文献 [18] 的工作，设计了能够处理连读误识样本的训练任务，连读误识可以被认为是插入和删除操作混合作用的结果。这让模型不仅适用于混淆词对，而且对由连读误识样本生成的混淆组块都有更强的适应能力。主要贡献包括：1) 提出以发音混淆组块作为预测目标，构建语言模型的训练方法。让预训练模型对组块更有效地进行表示；2) 设计抗发音混淆错误的意图识别模型，在构建的 ASR 口语语言理解数据集上精度超越多个基线模型，验证了该方法的有效性；3) 在基准口语理解数据集中，意图模型仍能保持有竞争力的精度，验证了该方法的泛化能力。

1　ASR 鲁棒的预训练语言模型

1.1　数据描述和问题定义

在应用意图识别模型的场合，常会获得两种类型的文本数据：用户直接输入的正确文本和由 ASR 转录用户语音得到的文本。为方便表述，以下分别称这两种数据为正确文本$\mathcal{D}_{\text{man}}$和误识文本$\mathcal{D}_{\text{asr}}$。后者常含有语音识别错误，甚至会影响标注人员对会话语义的理解。为了节约标注成本、提高标注质量，标注人员会优先标注正确文本用于模型训练。因此，本文的目标是帮助只用正确文本训练的意图识别模型，应对正确文本和错误文本测试时仍保持高精度。

算法包含 3 个步骤，1) 在通用数据上训练预训练语言模型，得到基础模型。2) 在无标注领域数据上，微调基础模型得到组块混淆感知语言模型 (text-chunk confusion-aware language model, TCLM)。每条无标注数据由转录相同语音的正确和误识文本对构成。实际应用中，这种数据可以复用 ASR 的训练数据，因为它包含语音和正确文本。3) 使用带有意图标签的正确文本，微调 TCLM 得到意图识别模型。

1.2　预训练语言模型的微调

文献 [20] 提出统一语言模型微调方法 ULMFit，它认为，尽管通用语言模型能够对常见语言有很强的表征能力，但是由于其训练时所用语料的数据分布与下游任务数据存在不同，需要使用任务数据微调通用语言模型，提高模型在任务上的精度。遵循该框架，本文采用掩码语言模型的训练方法，在领域数据上微调预训练语言模型 BERT[21]。即在给定文本$u=\{w_0,w_1,\cdots,w_{|u|}\}$ (w是词汇，$u\in\{\mathcal{D}_{\text{man}},\mathcal{D}_{\text{asr}}\}$)，掩码标记$w_{\text{mask}}$和被掩码词汇的索引集合$\mathcal{I}$时，定义掩码语言模型的损失函数为

$$\mathcal{L}_{mlm}=-\frac{1}{|J|}\sum_{i\in J}\ln p(w_i|w_{\text{mask}},w_{\neq i})$$

式中：$p(w_i|w_{\text{mask}},w_{\neq i})$是使用掩码标记替换目标词后的文本对目标词的预测概率。

这种方法能够加强语言模型分别对正确和错误文本的建模能力。比如，对于掩码文本“please w_{mask} song X to my playlist.”，模型有能力预测目标词大概率为 add（正确文本）和 at（误识文本）。但是由于缺乏显示指导，它不能利用两个目标词发音相似的这种客观关系。我们提出的 TCLM 方法，能克服这个缺点。

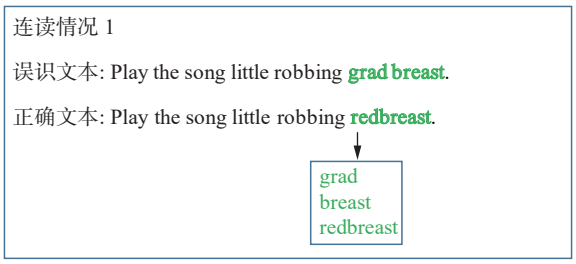
1.3　声学组块混淆感知语言模型微调技术

记文本u中t时刻的词是w_t^u, 从t到$t+\sigma$时刻的词序列是$s_t^u=\{w_t^u,w_{t+1}^u,\cdots,w_{t+\sigma}^u\}$。假设存在文本对$u_1$和$u_2$, 其中一个是正确文本, 另一个是误识文本。$w_{t_1}^{u_1}$和$s_{t_2}^{u_2}$在音素构成上具有较强相似性, 致使两者发生混淆。定义$\left(w_{t_1}^{u_1},s_{t_2}^{u_2}\right)$构成一个声学组块混淆实例, 其中$\boldsymbol{e}_t^u$表示词汇$w_t^u$在所处上下文环境中的词嵌入向量。声学组块混淆损失项为

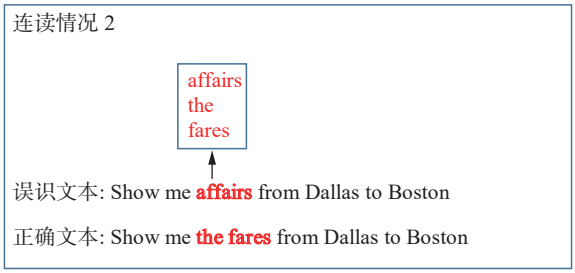
$$\mathcal{L}_{\text{accl}}=-\frac{1}{|\sigma+1|}\left(\ln p\left(s_{t_2}^{u_2}|\boldsymbol{e}_{t_1}^{u_1}\right)+\ln p\left(w_{t_1}^{u_1}|\boldsymbol{e}_{t_1}^{u_1}\right)\right)$$

式中, $\ln p\left(s_{t_2}^{u_2}|\boldsymbol{e}_{t_1}^{u_1}\right)=\sum_{i=t_2}^{t_2+\sigma}\ln p\left(w_i^{u_2}|\boldsymbol{e}_{t_1}^{u_1}\right)$。它在给定$\boldsymbol{e}_{t_1}^{u_1}$时, 把$s_{t_2}^{u_2}$中所有包含的词, 看作多分类问题的标签, 通过这种方式拉近组块声学混淆实例中两元素之间的距离。采用 Kullback-Leibler 距离来度量预测标签与真实标签之间的距离, 训练模型。

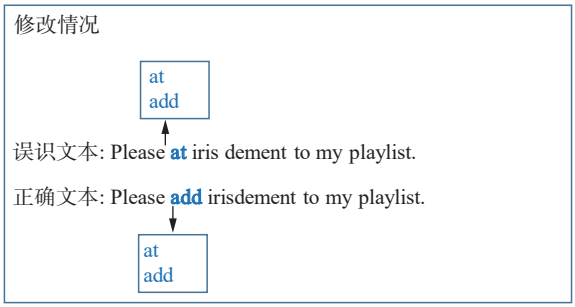
举例来讲, 在图 1 连读情况 1 中, 正确的文本 redbreast 被误识别成了两个单词 grad 和 breast。构建的声学混淆实例为(redbreast,{grad, breast})。定义的模型训练目标是使 redbreast 预测 grad、breast 和 redbreast 的概率最大化。在连读情况 2 中, the fares 被误识别成 affairs, 就可以构建声学混淆实例(affairs,{the,fares}), 而在修改情况中, 会产生实例(at,{add})和(add,{at})。后两种情况训练任务与情况 1 类似, 此处不再赘述。



(a) 连读情况 1



(b) 连读情况 2



(c) 连读情况 3

图 1　$\mathcal{L}_{accl}$应用到 ASR 3 种错误情况中的图例

Fig. 1　$\mathcal{L}_{accl}$ for 3 types of ASR errors

1.4　联合微调技术

在微调阶段, 联合优化目标函数包括掩码语言模型损失和声学组块混淆损失:

$$\mathcal{L}=\mathcal{L}_{\text{mlm}}+\beta\mathcal{L}_{\text{accl}}$$

式中: 超参数β用于平衡两个损失函数的贡献。该过程使模型同时考虑目标域语言特点和声学相似性。

1.5　口语语言理解

意图模型结构如图 2 所示, 它使用 1.2~1.4 节中预训练语言模型对文本进行上下文编码, 得到词嵌入向量, 经过最大池化操作、线性变换后, 采用 softmax 操作预测每个意图的概率, 选用交叉熵损失函数指导训练。

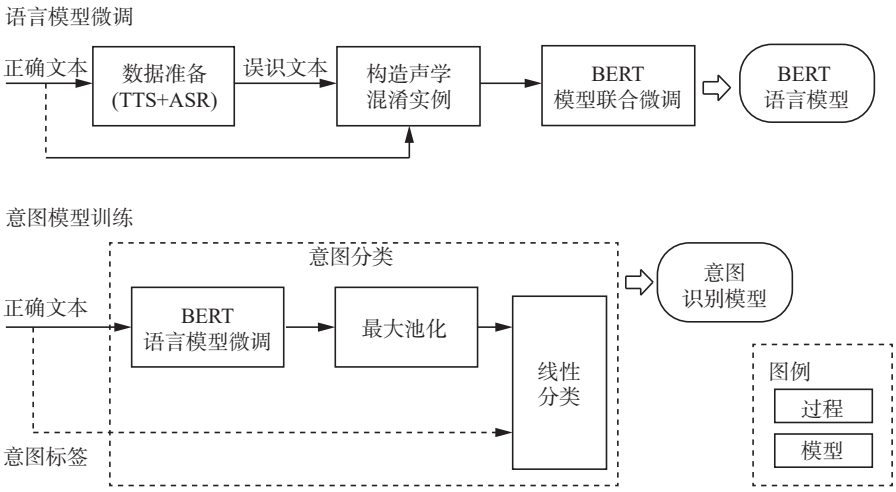


图 2　模型训练示意

Fig. 2　Model training schematic

2　实验与结果分析

2.1　实验设置

实验中所使用的 SLU 数据集如下：

1) Snips[22] 是用于对 SLU 系统进行基准测试的数据集。该数据集不包含语音录音，而文献 [18] 为进行意图实验，通过语音合成系统生成了语音数据，再使用 Kaldi[23] 提供的 ASR 模型生成的对应转录文本。为了公平比较，本文采用完全相同的数据做训练和测试。

2) ATIS[24] 是一个广泛使用的语言理解的基准数据集。此数据集与 Snips 相似，仅包含用户在使用航班信息系统过程中对应的正确文本，无语音数据及 ASR 转录文本。使用自研商用的 TTS 和 ASR 服务，先合成语音然后得到误识文本。

3) TOP[25] 是由 Facebook 推出的单轮对话数据集，会话集中在导航、事件以及导航到事件领域。同样因为此数据集不含 ASR 转录文本，需要使用 TTS 和 ASR 服务得到误识文本。在实验中仅使用了原数据集中的训练集和测试集。表 1 给出了上述数据集的统计信息，并在随机抽样的样本中统计连读错误占比。

表 1　数据集信息

Table 1　Statistic information of dataset

名称	训练规模	测试规模	意图数目	字错误率/%	连读错误占比/%
Snips	13 084	700	7	45.56	25.86
ATIS	4 778	893	21	38.16	19.23
TOP	31 279	9 042	25	27.32	42.42

我们使用 Kaldi 提供的自动文本对齐工具 align-text，对齐正确文本和误识文本，得到对齐文本对列表 confs 后，并采用如下描述，构造声学组块混淆实例。

算法　构造声学混淆实例

输入　对齐文本对列表 confs，其中列表元素是正确词 ref 和与之对齐的 ASR 转录词 hyp。当由于多、少词而导致它们不能对齐时，会用符号 <eps>进行补齐。

文本对列表样例：flights flights; that that; arrive <eps>; in arriving; baltimore baltimore

输出：声学组块混淆实例。

实例样例：(“arriving”, [“arrive, in”])

1) instances = []

2) cache = []

3) for ref, hyp in confs:

4) if hyp == “<eps>”:

5) cache.append(ref)

6) elif hyp != ref:

7) cache.append(ref)

8) instances.append((hyp, cache))

9) cache = []

10) return instances

实验所使用的数据设置如下：

1) 设置 1：采用正确文本分别做训练和测试。

2) 设置 2：采用误识文本分别做训练和测试。

3) 设置 3：采用正确文本做训练，误识文本做测试。本文算法主要关注在该设置下，算法的精度表现。

2.2　模型和训练细节

模型采用了 12 层 BERT(bidirectional encoder representation from transformers) 的预训练权重，输出层大小为 768 维。语言模型的学习率设置为 5×10^{-5}，最大迭代训练 20 轮。通过实验选择 σ 为 2。在微调阶段，β 设置为 1。batch 设置为 32，使用 AdamW[26] 作为优化器。分类模型的学习率设置为 1×10^{-5}，最大迭代训练 8 轮。

2.3　比较基线

选择如下方法进行实验比较：

1) Oracle:以 BERT 为主干网络的意图分类方法。但是与其他方法不同，该方法还会使用设置 2，注意只有此方法会用到该设置方式。

2) BERT：使用 BERT 作为主干网络，使用预训练模型做权重的初始化；

3) SpokenVec[18]：选择 ELMo[16] 作为主干网络，并用发音混淆的词汇对模型进行微调；

4) UMLFit：使用 BERT 作为主干网络，使用 1.2 节的方法对语言模型进行微调；

2.4　实验结果

2.4.1　σ 参数实验

表 2 所示在 3 个数据集上 σ 参数的实验结果。当 σ=0 时，声学组块退化为词汇对，是算法的最简化形式。当 σ=1 时，因为适当地扩大声学组块中元素的个数，所以能够提高算法的准确率。当 σ=2 时，将影响算法精度。这是由于参数值越大，在组块中包含的词汇越多，两条转录文本之间的句长差别越大，自动对齐工具的精度越低，所构造的声学组块和混淆实例也越不准确。

表 2　参数 σ 对算法精度的影响

Table 2　Effect of parameter σ on accuracy　%

σ	Snips	ATIS	TOP
0	92.55	**94.64**	90.18
1	**93.40**	94.31	**91.35**
2	92.27	94.20	90.83
3	92.97	93.97	90.95

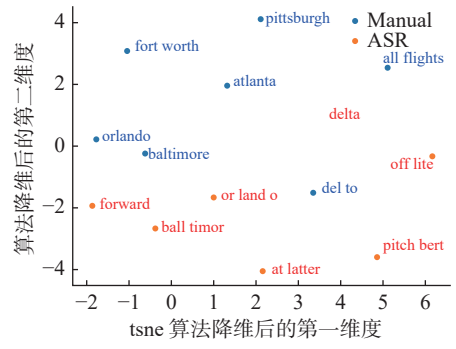
2.4.2　预训练模型定性分析

为了直观地观察预训练模型对声学组块混淆实例的表达和鉴别能力。在表 3 中，从 AITS 数据集中随机选择多对声学组块混淆实例及其上下文，分别使用 BERT 和 TCLM 预训练模型提取混淆实例在各自上下文条件下的嵌入向量表示。其中表格每行中的两个句子分别是正确文本和识别文本。图 3 显示了使用 t-SNE[27] 对嵌入向量进行二维投影的结果。可以看出使用原始的 BERT 模型，发音相似的实例之间分布散乱。而使用 TCLM 模型则能够合理地将发音相似的混淆实例聚合在一起，如“orlando”和“or land o”、“baltimore”和“ball timor”等，同时发音不同的实例之间的距离仍然得以保持，从图中可以清晰发现两者的差别。

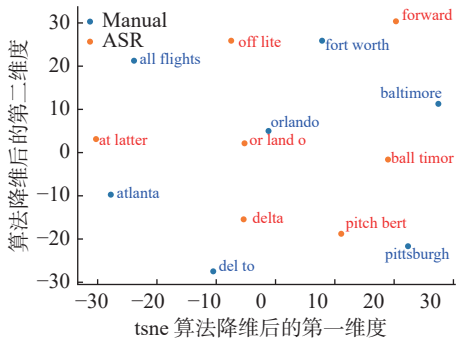
表 3　随机选择的声学组块混淆实例

Table 3　Examples of text-chunks

序号	正确文本	误识文本	混淆实例
1	does delta fly from atlanta to san francisco	does del to fly from atlanta to san Francisco	delta del to
2	show me all flights arriving to denver from boston	show me off lite arriving to denver from boston	all flights off lite
3	a flight from washington to fort worth	a flight from washington to forward	fort worth forward
4	from pittsburgh to baltimore	from pitch berta ball timor	pittsburgh pitch bert
5	give me flights from atlanta to baltimore	give me flights from at latter to baltimore	atlanta at latter
6	give me flights from atlanta to baltimore	give me flights from atlanta to ball timor	baltimore ball timor
7	i want to fly from milwaukee to orlando	i want to fly from milwaukee to or land o	orlando or land o



(a) BERT



(b) TCLM

图 3　嵌入表示的 2D 可视化

Fig. 3　2D visualization of embeddings

2.4.3　SLU 结果分析

表 4 所示是在不同设置下的实验结果。每个数据集的最佳结果以粗体标记。

从可以得出，实验结果呈现了如下现象：

1）Oracle 方法在设置 2 的精度显著低于设置 1。说明与以正确文本为处理对象的意图识别任务相比，误识文本的意图识别任务会更加困难。即使采用误识文本训练模型，仍然难以取得高测试精度。

2）BERT 方法在设置 3 的精度进一步降低。因为正确文本和误识文本之间特征分布存在较大差异。朴素地使用正确文本训练的模型，当遭遇误识文本时，可能严重失效。

3）SpokenVec 方法合理地利用了混淆词汇的特性，改进模型训练任务。尽管 Spoken Vec 所用的主干网络 ELMo，通常被认为是弱于 BERT 的，但在 3 个数据集上的表现都优于 BERT 方法。

4）UMLFIT 方法使用正确和错误文本训练语言模型，提升主干网络 BERT 对错误样本的建模能力。精度有较大提升，甚至超过了 SpokenVec 方法，是很强的参考基线。因为它排除了 $\mathcal{L}_{\text{accl}}$ 损失项，可视为本文方法的消融实验结果。

5）在设置 3，本文方法优于所有基线方法。相比于同样利用混淆词汇的 SpokenVec 方法，该方法在 3 个数据集上的绝对精度分别提升 4.88%、3.02% 和 2.39%，验证了方法的有效性。这得益于使用了更强的主干网络，以及使用声学组块混淆实例的预训练方法。有趣的是，该方法的精度有时还会超过 Oracle 方法，猜测是由于从混淆实例中，通过关联对比到了更多知识。

6）另外，即使在设置 1，本文方法也能够保持有竞争力的精度。尽管在模型预训练阶段引入误识文本，但在意图模型训练阶段使用的都是正确文本，这会让模型保持对正确文本的处理能力；尽管 Snips 与其他数据集的误识文本是由不同的

ASR 引擎得到的，但是本文方法都有一致的表现，说明该方法对 ASR 算法没有特殊要求，是一种普适的方法，在工业场景，容易与给定 ASR 引擎对接适配。

表 4　不同数据集下意图检测任务的准确率

Table 4　Accuracy of Intent detection　%

设置 1:　设置 2:　设置 3:

方法	Snips 数据集		ATIS 数据集		TOP 数据集	
	正确文本	误识文本	正确文本	误识文本	正确文本	误识文本
Oracle	97.86	92.26	97.32	95.65	95.01	92.63
BERT	97.86	77.82	97.32	89.15	95.01	86.14
SpokenVec	97.01	88.52	95.42	91.29	93.55	88.96
UMLFIT	**98.43**	90.55	**97.43**	92.52	94.72	89.96
本文方法	98.43	93.40	97.32	94.31	94.68	91.35

但是，通过对结果数据的观察，本文方法也有一定的局限性。因为需要借助对齐算法生成混淆实例，对齐算法的优劣会对算法产生影响。另外对于在预训练过程中没有见到的混淆实例，意图模型的表现欠佳，后续有可能通过扩大预训练的数据规模缓解这个问题。

3　结束语

本文提出了一种新的声学组块混淆感知语言模型学习方法，能够从声学组块混淆实例中，学习到对 ASR 连读、删除误识鲁棒的词嵌入表示。在多个意图分类数据集上的实验表明，所学到的词嵌入表示应用到该任务时，能够提高模型对误识文本意图识别的鲁棒性。在未来，我们还会研究如何将该学习方法应用在其他 SLU 的槽位/值识别任务上，探究算法的鲁棒性表现。

参考文献：

[1] 程高峰, 颜永红. 多语言语音识别声学模型建模方法最新进展 [J]. 计算机科学, 2022, 49(1): 47–52.
CHENG Gaofeng, YAN Yonghong. Latest development of multilingual speech recognition acoustic model modeling methods[J]. Computer science, 2022, 49(1): 47–52.

[2] 赵宁, 徐俊利, 徐洋航, 等. 客户来电意图识别研究 [J]. 中文信息学报, 2021, 35(3): 125–133.
ZHAO Ning, XU Junli, XU Yanghang, et al. Intention detection of customer's call[J]. Journal of Chinese information processing, 2021, 35(3): 125–133.

[3] 吕坤儒, 吴春国, 梁艳春, 等. 融合语言模型的端到端中文语音识别算法 [J]. 电子学报, 2021, 49(11): 2177–2185.
LYU Kunru, WU Chunguo, LIANG Yanchun, et al. An end-to-end Chinese speech recognition algorithm integrating language model[J]. Acta electronica sinica, 2021, 49(11): 2177–2185.

[4] 徐扬, 王建成, 刘启元, 等. 基于上下文信息的口语意图检测方法 [J]. 计算机科学, 2020, 47(1): 205–211.
XU Yang, WANG Jiancheng, LIU Qiyuan, et al. Intention detection in spoken language based on context information[J]. Computer science, 2020, 47(1): 205–211.

[5] 李蕾, 周延泉, 钟义信. 基于语用的自然语言处理研究与应用初探 [J]. 智能系统学报, 2006, 1(2): 1–6.
LI Lei, ZHOU Yanquan, ZHONG Yixin. Pragmatic information based NLP research and application[J]. CAAI transactions on intelligent systems, 2006, 1(2): 1–6.

[6] SERDYUK D, WANG Yongqiang, FUEGEN C, et al. Towards end-to-end spoken language understanding[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary: IEEE, 2018: 5754−5758.

[7] HAGHANI P, NARAYANAN A, BACCHIANI M, et al. From audio to semantics: approaches to end-to-end spoken language understanding[C]//2018 IEEE Spoken Language Technology Workshop. Athens: IEEE, 2018: 720−726.

[8] LUGOSCH L, RAVANELLI M, IGNOTO P, et al. Speech model pre-training for end-to-end spoken language understanding[C]//20th Annual Conference of the International Speech Communication Association. Graz: ISCA, 2019: 814−818.

[9] HUANG Yinghui, KUO H K, THOMAS S, et al. Leveraging unpaired text data for training end-to-end speech-to-intent systems[C]//ICASSP 2020−2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020: 7984−7988.

[10] KUO H K J, TÜSKE Z, THOMAS S, et al. End-to-end spoken language understanding without full transcripts[C]//21st Annual Conference of the International Speech Communication Association, Shanghai: ISCA, 2020: 906−910.

[11] SUNDARARAMAN M N, KUMAR A, VEPA J. Phoneme-BERT: joint language modelling of phoneme sequence and ASR transcript[EB/OL]. (2021−02−01) [2022−09−12].https://arxiv.org/abs/2102.00804.

[12] ŠVEC J, ŠMÍDL L, IRCING P. Hierarchical discriminative model for spoken language understanding[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver: IEEE, 2013: 8322−8326.

[13] ŠVEC J, CHÝLEK A, ŠMÍDL L, et al. A study of differ-

ent weighting schemes for spoken language understanding based on convolutional neural networks[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai: IEEE, 2016: 6065−6069.

[14] LADHAK F, GANDHE A, DREYER M, et al. LatticeRnn: recurrent neural networks over lattices[C]//17th Annual Conference of the International Speech Communication Association. San Francisco: ISCA, 2016: 695−699.

[15] HUANG Chaowei, CHEN Yunnung. Learning spoken language representations with neural lattice language modeling[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 3764−3769.

[16] WENG Yue, MIRYALA S S, KHATRI C, et al. Joint contextual modeling for ASR correction and language understanding[C]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020: 6349−6353.

[17] MASUMURA R, IJIMA Y, ASAMI T, et al. Neural confnet classification: fully neural network based spoken utterance classification using word confusion networks[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary: IEEE, 2018: 6039−6043.

[18] HUANG Chaowei, CHEN Yunnung. Learning asr-robust contextualized embeddings for spoken language understanding[C]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020: 8009−8013.

[19] NAMAZIFAR M, TUR G, HAKKANI-TÜR D. Warped language models for noise robust language understanding[C]//2021 IEEE Spoken Language Technology Workshop. Shenzhen: IEEE, 2021: 981−988.

[20] HOWARD J, RUDER S. Universal language model fine-tuning for text classification[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne: Association for Computational Linguistics, 2018: 328−339.

[21] DEVLIN J, CHANG Mingwei, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019: 4171−4186.

[22] COUCKE A, SAADE A, BALL A, et al. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces[EB/OL]. (2018−05−25)[2022−09−12].https://arxiv.org/abs/1805.10190.

[23] POVEY D, GHOSHAL A, BOULIANNE G, et al. The Kaldi speech recognition toolkit[C]//IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, 2011.

[24] HEMPHILL C T, GODFREY J J, DODDINGTON G R, et al. The ATIS spoken language systems pilot corpus[C]//Proceedings of the workshop on Speech and Natural Language−HLT'90. Hidden Valley: Association for Computational Linguistics, 1990: 96−101.

[25] GUPTA S, SHAH R, MOHIT M, et al. Semantic parsing for task oriented dialog using hierarchical representations[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018: 2787−2792.

[26] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[C]//7th International Conference on Learning Representations. New Orleans LA: ICLR, 2019.

[27] VAN DER MAATEN L, GEOFFREY H. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9: 2579–2605.

作者简介:

李荣军, 主任工程师, 主要研究方向为人机对话、语音识别。

郭秀焱, 高级工程师, 主要研究方向为知识图谱、人机对话、语音识别。

杨静远, 高级工程师, 主要研究方向为智能问答、任务型对话系统、语音纠错。