文章编号：1003-0077（2017）00-0000-00

结合特殊领域实体识别的远监督话语领域分类

何宇虹 黄沛杰 杜泽峰 刘威 朱建恺
（华南农业大学 数学与信息学院，广东省 广州市 510642）

摘要：近年来，基于注意力（attention）机制的循环神经网络在文本分类中表现出显著的性能。然而，当训练集数据有限时，测试集数据中许多领域实体指称项在训练集中处于低频，甚至从未出现，如，中文话语领域分类任务。本文提出结合特殊领域实体识别的远监督话语分类模型。首先，通过远监督（distant supervision）的方式获取数据集中的领域知识，显著地减少了人工操作；其次，利用特殊领域实体识别和本地构建的补充性知识库去补全远监督获取的领域知识，旨在为模型提供更加全面的领域知识；最后，对基于上下文的语义特征和知识特征这两种异构信息提出了细粒度拼接机制，在词级上融合了预训练词汇语义表达和领域知识表达，有效地提升了分类模型的性能。通过与研究进展的文本分类模型的对比实验表明，本文的模型在中文话语领域分类基准数据集的实验上取得了较高的正确率，特别是在知识敏感型领域，较研究进展方法具有显著的优势。

关键词：领域分类；外部知识；远监督；话语表达；神经分类器

中图分类号：TP391 **文献标识码：**A

Distant Supervision Based Utterance Domain Classification with Domain-Special NER

HE Yuhong, HUANG Peijie, DU Zefeng, LIU Wei and ZHU Jiankai

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

Abstract: Recently recurrent neural networks with an attention mechanism have achieved strong results on text classification. However, when the labeled training data is not substantial, such as in Chinese utterance domain classification (DC) task, many domain entity mentions have low frequency or are unseen in the training data, which presents a significant challenge. To address this issue, this paper proposes knowledge-based neural DC (K-NDC) models that incorporate domain knowledge from external sources into neural DC to enrich the representations of utterances. Firstly, decent entities and types are obtained by distant supervision from CN-Probase. Then domain-special named entity recognition (NER) and complement KB are exploited to further complement the knowledge coverage. Finally we design a novel mechanism for merging knowledge with utterance representations at fine-grained (Chinese word level). Experiments on the SMP-ECDT benchmark corpus show that comparing with the state of the art text classification models the proposed method achieves a strong performance, especially in knowledge-intensive domains.

Key words: domain classification; external knowledge; distant supervision; utterance representation; neural classifier

0 引言

口语智能助手在我们的日常生活中变得越来越重要，很多日常便携设备，如手机、智能手表、电脑都引入了智能助手应用程序[1]。智能助

收稿收稿日期：；**定稿日期**：=
基金项目：国家自然科学基金(71472068)；广东省大学生创新训练计划项目(201810564094, 201910564164)

手的关键组成部分是口语语言理解（spoken language understanding，SLU）模块，即机器通过理解用户特定的指令，有针对性地去执行任务[2]。这种对“特定指令”进行理解的第一步是，将接收到的用户的话语分类到特定的领域中，以便进行进一步处理。此过程称为话语领域分类（domain classification，DC）[3-4]。

随着深度学习技术的迅速发展，一系列深度神经网络（deep neural networks，DNNs）被应用于文本分类的任务上[3]。最近，循环神经网络（recurrent neural network，RNN）[5-7] 由于具有较好的随时间保存序列信息的能力而被广泛应用于文本分类任务。此外，有研究工作在 RNN 的基础上，引入了注意力（attention）机制，使得 RNN 能够选择性地关注序列中特定的信息，并且在分类正确率上取得了显著的效果。

尽管上述研究表明，现有的神经文本分类器在经验上表现良好，但从话语领域理解的角度来看，它们仍然存在局限性。首先，这些神经分类器的有效性依赖于大量的标注训练数据，即，当训练数据有限时，如，在中文话语领域分类任务（如本文实验采用的中文话语领域分类基准语料 SMP-ECDT）中，测试集中许多领域实体指称项（entity mention）在训练集中很低频，甚至没出现过，这就导致这些领域实体指称项在注意力机制中的权重是一个趋近 0 的数，从而造成了信息损失和模型性能下降。

其次，现有的神经文本分类器欠缺了对领域实体指称项分词不合理的考虑。分词是处理中文文本的一个必备基本操作，现有的流行分词工具对领域实体指称项分词都存在局限性，这限制了领域实体指称项语义表达的准确性。例如，表 1 句子 2 的歌曲名实体指称项“一路上有你” 对应的分词是“一路上”、“有”和“你”， 显然他们之间的语义是不等价的，即在领域分类任务中，领域实体指称项分词的组合无法表达实体指称项的原始语义[8]，从而导致错误的语义被传递。

此外，当用户对智能助手说出一句全知识话语（话语仅包含领域实体指称项）时，其意图在于给出面向任务的命令，而不是闲聊。如表 1 句子 3，“我只在乎你” 旨在播放“我只在乎你”这首歌，即意图（领域分类）是“音乐”，而不是闲聊。

表 1 SMP-ECDT 数据集示例

序号	话语	领域
1	打开优酷网	网站
2	张学友的一路上有你	音乐
3	我只在乎你	音乐
4	横看成岭侧成峰的下一句是什么？	诗词

为了解决这些局限性，本文提出了融合知识的神经话语领域分类模型（knowledge-based neural utterance domain classification，K-NDC），即利用外部领域知识来丰富话语表征学习。具体地，我们的目标是如何适当地利用知识库（knowledge bases，KBs）中的实体及其领域标签信息构建知识特征去增强语言的理解。

在 KBs 的背景下，远监督方式利用远程外部 KBs 中的实体及其领域标签信息去获取数据集的领域知识，能够降低昂贵的人力成本[9-12]。在本文中，我们首先依赖于通用知识库 CN-Probase，通过远监督[13]初步获取数据集的领域知识。由于通用知识库应用于特殊领域的任务会造成一些实体及其领域标签信息缺失，我们运用特殊领域实体识别[14]去抓取丢失的实体指称项，并利用可靠的额外知识来源构建本地知识库去补充缺失的领域标签信息，旨在为模型提供更加全面的领域知识。

最后，基于获取的领域知识，我们研究了如何将知识融入到 NDC 模型中。许多已有研究采用了粗粒度拼接机制，利用句子级别的知识表达来丰富话语表达[15-16]。尽管该机制在一定程度上能够将领域知识作为背景特征提供给分类器，但它忽略了知识对话语中每个词语的影响。为此，本文设计了一个细粒度拼接机制，在词级别上，构建词汇语义表达和领域知识表达之间的匹配，此机制在实验中表现了具有良好的性能。本文的主要贡献如下：

(1) 提出了融合知识的神经话语领域分类模（K-NDC）。首先采用远监督初步获取数据集的领域知识，然后利用特殊领域实体识别和可靠的额外知识来源构建的本地 KB，为模型提供更加全面的领域知识，进而将得到的知识表达结合到神经网络分类器中进行话语领域分类。

(2) 设计了一种细粒度拼接机制，在词级上融合了预训练词汇语义表达和领域知识表达，使得词语的向量表达融入了知识因素，提升了以其作为输入的神经分类器的性能。

(3) 本文提出的方法，在中文话语领域分类基准语料 SMP-ECDT 上，取得了优于研究进展的神经分类模型以及粗粒度拼接的知识增强基线模型的分类性能，尤其在知识敏感型领域有显著的提升。此外，实验还验证了本文的知识库增补方案的良好效果。

1 相关工作

话语领域分类属于短文本分类，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]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（Kenter and Rijke），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[20]或者使用神经网络训练词向量的方法增加语义特征。在话语领域分类方面，早期的领域分类还结合口语话语的特点，采用了一些较为复杂的人工特征，如语法信息、韵律信息、词汇信息等[21-23]。

近年来，深度学习在自然语言处理（natural language processing, NLP）中受到关注[24]，主流的应用包括深度信念网络（deep belief network，DBN）[24]、CNN[25-26]和 RNN[4]，尤其是 RNN 中最常用的 LSTM (long short-term memory network)[27-30]。随后，注意力机制被引入到了 NLP 中，实验证明其善于在文本分类任务中抽取文本的含义，例如意图检测[31]、领域分类[32]和文档分类[33]等。本文的神经分类模型选用了短文本分类（尤其是话语分类）研究进展中较为有代表性的带注意力机制的双向 LSTM（BiLSTM）[31]，研究如何在其基础上有效地融入外部知识。

在基于知识的 NLU 方面，语言知识[34-35]或知识库（knowledge bases，KBs）[15-16, 36]被视为先验知识用于辅助语言理解。在本文中，我们的目标是恰当地利用 KBs，对话语进行实体识别和实体链接获取知识特征来强化 NDC 模型。为了融合知识信息和神经网络模型，大多数已有研究基于生成一个编码器的嵌入层的想法对外部知识进行建模，知识特征和话语的文本特征之间采用了句子级别的粗粒度拼接机制[15-16]。虽然该机制在一定程度上也能够将知识信息以背景特征的形式融入神经分类器，但它忽略了知识信息在传播过程中对句子中每个词语权重分配的影响。与上述研究不同，本文采用了细粒度拼接机制，在词级别上构建词汇语义特征和知识特征之间的匹配。

最后是关于外部知识。远监督是一种利用大型远端 KB 中实体标签信息或实体关系信息来代替大量人工操作的技术，近年来被广泛应用于关系抽取（relation extraction）[9]和实体标签（entity typing）[10-11]的研究工作中。本文运用远监督获取实体及其领域标签，具体一般的流程如下[11]：（1）识别句子中的实体指称项；（2）链接被识别的实体指称项到知识库中的实体[12, 37]；（3）将实体在知识库的领域标签分配给句子中的实体指称项[12]。然而，远监督通常是适用于通用领域实体标签获取，当应用于特殊领域的任务中（如本限定领域的领域分类任务），会造成一些实体及其领域标签信息缺失。本文通过进一步采用特殊领域实体识别和构建补充性的本地 KB，辅助远监督提供更加全面的外部知识。

2 本文的方法

2.1 模型结构

本文提出的融合外部知识的神经话语领域分类模型 K-NDC 模型结构如图 1 所示。

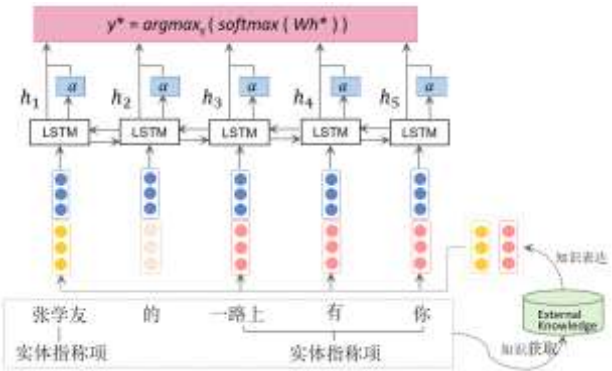

图 1 融合外部知识的神经话语领域分类模型 K-NDC

模型主要由两个模块组成：外部知识模块和融合知识的神经分类器。我们将在 2.2 和 2.3 小节分别具体介绍这两个部分。给定一个输入话语，首先通过外部知识模块获取话语中的知识实体指称项以及相应的实体标签，形成的知识表达通过词级细粒度拼接方式扩展了话语的词汇向量表达，进而输入到带注意力的 BiLSTM 得到预测的领域分类结果。

2.2 外部知识

我们依赖于 CN-Probase[13]，通过远监督技术初步获取句子中的实体指称项以及实体指称项对应的领域标签信息。CN-Probase 是一个流行的通用中文实体标签库[38]，提供了大量的“实体-标签”事实对。然而，在特定领域的分类任务中，它存在实体缺失和标签不全的问题。为了解决通用的 CN-Probase 在特定领域的泛化问题，我们引入特殊领域实体识别和本地知识库，旨在为模型提供更全面的实体和领域标签信息。特别地，由于 CN-Probase 是综合全面的，覆盖了许多领域，本文只保留了与当前任务相关领域的标签。同时，在输出知识特征表达时，我们进一步合并过于细致的标签，如“诗名”、“诗人”、“诗句”被合并成同一领域标签“诗词”。图 2 为本文获取外部知识的技术架构。

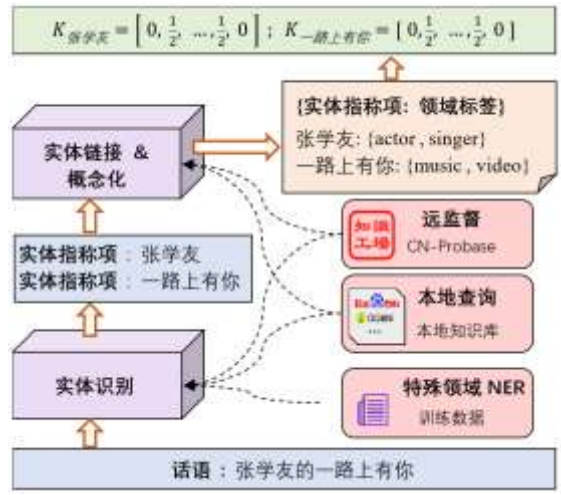

图 2 获取外部知识的技术架构

2.2.1 特殊领域实体识别

远程的通用领域实体识别在特定领域的分类任务中出现丢失实体及其领域标签的情况，从而限制了分类模型的性能。并且，当数据集中的实体数量越多，这种丢失越严重，性能下降就越明显（如后面实验中的知识敏感型领域）。为了给分类模型提供更加充分的实体及其领域标签，我们在 SMP-ECDT 的训练集上人工标注出句子中的实体指称项和标签，然后在此基础上，采用 Lattice LSTM[14]训练特殊领域实体识别模型。为了保证实体领域标签的完整性，我们仅使用实体识别模型结果的实体指称项（放弃实体指称项的标签），进而采用远监督和本地知识库检索相结合的方式去链接实体和实体指称项，并获取实体的标签信息。

2.2.2 本体补充性知识库

由于语料有限，尽管引入了特殊领域实体识别，依然存在一些实体指称项因为太特殊而无法被准确识别出来，如，在“横看成岭侧成峰的下一句是什么？”中，不管是 CN-Probase 还是特殊领域实体识别，都无法识别出“横看成岭侧成峰”这一实体指称项。为了保证实体识别的覆盖率，本文利用可信的外部知识源，抓取并过滤获得特定领域的知识，构建了补充性的本地知识库。例如，在百度百科上抓取古诗、彩票、天气等领域的知识；在 QQ 音乐中抓取了歌手、歌曲等领域知识。

值得注意的是，出于可信度的考虑，我们放弃了字数太少的知识，如，少于 4 个字的歌曲名。

2.3 融合知识的神经分类器

2.3.1 基于带注意力的 BiLSTM 的话语领域分类

BiLSTM[39]是 RNN 的一种变形，使用正向和反向 LSTM 来处理序列 $X=[w_1,\dots,w_n]$。正向 LSTM 从左往右处理序列，产生正向隐藏状态 $\overrightarrow{h_t}=\overrightarrow{LSTM}\left(w_t,\overrightarrow{h_{t-1}}\right)$，反向 LSTM 从右往左处理序列，产生反向隐藏状态 $\overleftarrow{h_t}=\overleftarrow{LSTM}\left(w_t,\overleftarrow{h_{t-1}}\right)$。然后拼接正反向 LSTM 得到的隐藏状态 $h_t=[\overrightarrow{h_t}\oplus\overleftarrow{h_t}]$，得到序列的最终隐层状态：$h=[h_1,\dots,h_n]$。本文的基础神经分类器采用了话语分类研究进展中较为有代表性的带注意力机制的 BiLSTM[31]。

Liu 和 Lane[31]模型中的注意力机制是一种软注意力（soft attention）机制，首先是通过得分函数为每个隐藏状态分配注意力权重，然后计算一个注意力权重对齐的输入状态，最终得到输出分布，如式（1）-（3）所示：

$$s_i=score\left(h_i\right) \tag{1}$$

$$\alpha_i=\frac{\exp(s_i)}{\sum_i^n\exp(s_i)} \tag{2}$$

$$h^*=\sum_i^n\alpha_ih_i \tag{3}$$

其中 $score$ 为前馈神经网络，权重 $(\alpha_1,\dots,\alpha_n)$用于计算 BiLSTM 输出 $(h_1,\dots,h_n)$ 的加权和 h^*。接着，带权重的输出向量 h^* 作为序列的最终语义表达被送入 softmax 层，得到序列所有领域类别 $Y=[y_1,\dots,y_s]$ 的预测概率分布，如式（4）所示，其中 W 是权重矩阵。

$$P(Y\mid X)=softmax\left(Wh^*\right) \tag{4}$$

其中，$\text{softmax}\left(z_i\right)=\frac{e^{z_i}}{\sum_j^n e^{z_j}}$。

最后，取概率最大的领域类别作为当前序列（话语）的领域分类预测结果 y^*：

$$y^*=argmax_y\left(P(Y\mid X)\right) \tag{5}$$

2.3.2 细粒度词级拼接机制

句子基于上下文的语义特征和来自于外部知识的知识特征是两种异构信息，常用的融合两种异构信息的方法是以粗粒度（coarse-grain）的方式在神经分类器最后的 softmax 层之前拼接[15-16]。在领域分类任务上，这种拼接方式在一定程度上能够将领域知识作为背景特征提供给分类器。然而这种方式忽略了知识对句子中每个分词的影响，特别地，我们认为句子中词汇的权重仅考虑上下文因素是存在不足的。事实上，人在听到一句话时，关注的不仅仅是上下文，还有话语中的关键知识部分。例如，“对酒当歌人生几何的下一句是什么，你知道吗？”，知识词权重的考虑因素不应该等价于非知识词权重的考虑因素。因此，从用户意图表达学习的角度看，仅仅把知识特征作为背景信息会限制模型的表达能力，为了我们提出采用细粒度（fined-grain）的语义特征和知识特征融合方式。

（1）知识特征表达

对于给定的长度为 n 的话语 $U=[w_1,\dots,w_n]$，我们从外部知识模块获取所有的实体指称项及其领域标签。给定实体指称项 m 和对应的领域标签 t，定义其知识表达为 q 维向量 $K_m=[r_1,\dots,r_q]$，其中，q 表示整个数据集的领域标签个数，即 K_m 的每一维代表语料中的一个领域标签，计算方式如式（6）所示：

$$r_j=\frac{1}{|t_i|}*F(^*r_j,t_i) \tag{6}$$

其中，$|t_i|$ 是实体指称项 m 的领域标签个数，$^*r_j$指代 K_m 的第 j 维表示的领域标签，F 是指示函数，定义如式（7）所示。

$$\mathbb{I}(^*r_j,t_i)=\begin{cases}1, & if\ ^*r_j\in t_i\\ 0, & otherwise\end{cases} \tag{7}$$

例如，表1中的示例“张学友的一路上有你”，从外部知识模块获得实体指称项“张学友”和领域标签「音乐、视频」（由「歌手、演员」标签转换而来），则 $r_{music}=\frac{1}{2},r_{video}=\frac{1}{2}$，“张学友”的知识表达为 $K_{张学友}=[0,\dots,\frac{1}{2},\dots,\frac{1}{2},\dots,0]$。同理可

得，实体指称项“一路上有你”的领域标签是「音乐、视频」，则“一路上有你”知识表达为 $K_{一路上有你}=[0,\dots,\frac{1}{2},\dots,\frac{1}{2},\dots,0]$。

（2）词级拼接

细粒度拼接机制旨在以细粒度方式构建两个异构信息之间的匹配，建立词级别的语义特征和知识特征之间的联系，即句子中的每个分词都有其对应的知识表达。在细粒度拼接机制中，我们对每个分词的预训练词向量和对应知识表达构建匹配，如式（8）所示：

$$U^{*}=f_{match}(U,K)=[w_i \oplus K_i] \quad (8)$$

具体地，若子序列 $x=[w_i,\dots,w_{i+a}]$ 构成一个实体指称项 m，则子序列 x 中的分词具有相同的指向实体指称项 m 的知识表达，即 $K_i=\cdots=K_{i+a}=K_m$。而对于话语中的非知识分词，其对应的知识表达是维度为 q 的零向量。

通过细粒度词级拼接机制，话语的词汇表达进一步扩展为词向量表达和知识表达的组合，使得词汇的语义表达融入了知识因素。扩展后的词向量表达送入 2.3.1 小节的带注意力机制的 BiLSTM 模型，得到预测领域分类结果 y^{*}。

3 实验

3.1 数据集

本文选用中文口语领域分类基准语料 SMP-ECDT[40]作为实验数据。该语料由哈尔滨工业大学社会计算与信息检索研究中心（哈工大 SCIR）和科大讯飞股份有限公司（iFLYTEK）提供。具体上采用的是 2018 年第二届中文人机对话技术评测任务一“中文话语领域分类”中的数据集。该语料包含任务型垂直领域（如查询机票、酒店、公交车等）、知识型问答以及闲聊等共 31 个领域。训练集包含 3736 条语料，测试集包含 4528 条语料，均为单轮短文本话语。

本文用于训练 word2vector 的数据是由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）。本文使用了该数据集的一个 10G 的子集作为词向量的训练语料。

3.2 实验设置及评价方法

本文采用 Jieba 分词、C-BOW 预训练的词向量[8]分别对数据进行处理，OOV 词则随机初始化。实验中模型训练的参数调节均采用 k-折（本文采用 10 折）交叉验证。实验中调参组合如下：使用 RMSProp 优化器，学习率设为 0.001，词向量维度为 400，batch-size 为 25。为了防止过拟合，我们在训练过程中使用 dropout[41]。模型评测是对测试集做集成测试，并且采用多次独立实验，取平均值。

领域分类采用正确率作为评测方法。知识库增补方案（包括特殊领域的 NER 和本地补充性知识库）采用两种不同粒度的精确度、召回率和 F1 值作为评测方法[10-11]。定义话语集为 U，对于单句语料 u，t_u表示其标准实体指称项集，$\widehat{t_u}$表示模型预测得到的实体指称项集，则两种粒度的精确度（precision）和召回率（recall）计算方式如下：

- **Loose Macro**: 对于每句语料，精确度（precision）和召回率（recall）计算如式下：

$$precision=\frac{1}{|U|}\sum_{u\in U}\frac{|\widehat{t_u}\cap t_u|}{|\widehat{t_u}|} \quad (9)$$

$$recall=\frac{1}{|U|}\sum_{u\in U}\frac{|\widehat{t_u}\cap t_u|}{|t_u|} \quad (10)$$

- **Loose Micro**: 对于每句语料，精确度（precision）和召回率（recall）计算如下：：

$$precision=\frac{\sum_{u\in U}|\widehat{t_u}\cap t_u|}{\sum_{u\in U}|\widehat{t_u}|} \quad (11)$$

$$recall=\frac{\sum_{u\in U}|\widehat{t_u}\cap t_u|}{\sum_{u\in U}|t_u|} \quad (12)$$

3.3 对比方法

本文对多个研究进展模型进行实验，并与我们提出的 K-NDC 模型对比。

- BiLSTM：在处理序列数据时表现良好的性能，是文本分类的经典模型[29, 39]。
- Soft Attention：Liu 和 Lane[31]采用了基于 attention 机制的 BiLSTM 模型，在 ATIS 意图识别任务上取得良好的效果。我们在领域分类的

BiLSTM 模型上应用了他们论文里的软注意力机制。

- Multi-head Attention：Vaswani 等人[42]提出的 Multi-head 注意力模型，并应用在话语领域分类任务上，它的主要结构是“Transformer”。
- Hard Attention：我们在领域分类的 BiLSTM 模型上应用了 Shankar 等人[43]提出的硬注意力机制。

我们提出的两个 K-NDC 模型：

- K-CNDC：采用流行的粗粒度（Coarse-grain）拼接机制[15-16]，使用句子级别的知识表示来丰富话语表达。本文将该方法视为带知识的神经分类器的基线。
- K-FNDC：本文提出的最终方案，采用细粒度（Fined-grain）的拼接机制，建立词级别的语义表征和知识特征之间的联系。

3.4 实验结果和分析

3.4.1 与研究进展方法的对比

表 2 展示了我们的模型以及对比模型在 SMP-ECDT 数据集上的实验结果。实验结果显示，在话语领域分类任务上，本文提出的模型（K-NDC）的性能优于 BiLSTM、Soft Attention、Multi-head Att 和 Hard Att 四个研究进展神经分类模型，表明给模型增加领域知识，不管是以粗粒度的方式还是细粒度的方式，神经网络模型的分类能力得到提升，这表明领域知识能够帮助模型学习更高质量的语义特征。其次，相比知识基线模型 K-CNDC，K-FNDC 的分类正确率有了显著的提升。相比于最好的非知识基线神经分类器（Hard Attention）和 K-CNDC，K-FNDC 的领域分类正确率分别提高了 3.74%和 3.37%，表明本文提出的在细粒度层面上引入知识的机制，能为模型提供更加合理的语义信息，有助于提高神经分类器的分类性能。

表 2 本文提出方法与研究进展方法的对比

方法	Acc（%）
BiLSTM[29, 39]	76.97
Soft Attention[31]	78.23
Multi-head Attention[42]	76.74
Hard Attention[43]	78.84
K-CNDC	79.21
K-FNDC	82.58

3.4.2 知识库增补的效果

为了验证本文补全 CN-Probase 对模型在领域分类任务的影响，我们在不同知识库补全层级上进行模型训练（采用 K-FNDC），观察话语实体识别模块和模型在领域分类任务上的性能，结果如表 3 所示。“Full model”表示综合采用特殊领域 NER 和本地补充性 KB。

由表 3 可知，相比于只使用 CN-Probase，采用特殊领域 NER，以及进一步的知识库补全都带来 Ma-F1 和 Mi-F1 的一致提升，同时，领域分类效果也跟着提升，验证了本文知识库增补方案的良好效果。同时我们还注意到，仅通过 CN-Probase 远监督的召回率相对不足，这主要是因为它是面向通用应用的知识库。而在综合采用特殊领域 NER 和本地补充性 KB 之后，不论是 Macro 还是 Micro 的统计指标，R 值都有了显著提高，较为完善的实体识别确保了较为完整的知识表达融入到神经分类器中，使得模型的领域分类正确率达到了 82.58%，比仅采用 CN-Probase 远监督的神经分类器的分类正确率提高了 3.19%。

3.4.3 知识敏感型领域效果分析

为了进一步与已有模型对比，我们选取了 8 个知识敏感型领域（*music*, *app*, *datetime*, *poetry*, *radio*, *tvchannel*, *video* 和 *website*）。这些领域的话语中含有较多知识实体，如歌曲名、节日、电影名等。图 3 展示了不同神经 DC 模型在这 8 个知识敏感型领域的性能表现。

从图 3 可以看到，本文提出的 K-FNDC 模型在知识敏感型领域有优异的表现，比最好的非知识基线模型的分类正确率提升达到了 9.8%。此结果也同时表明，知识在话语领域分类任务中起着

重要的作用。

表 3　特殊领域实体识别和本地补充性 KB 在 NER 模块和 DC 任务的表现

外部知识方案	NER						DC
	Ma-P	Ma-R	Ma-F1	Mi-P	Mi-R	Mi-F1	Acc（%）
K-FNDC (CN-Probase)	0.678	0.414	0.513	0.551	0.410	0.470	79.39
K-FNDC (CN-Probase + domain-specific NER)	0.680	0.801	0.735	0.607	0.787	0.685	81.54
K-FNDC (Full model)	0.702	0.938	0.803	0.620	0.930	0.744	82.58

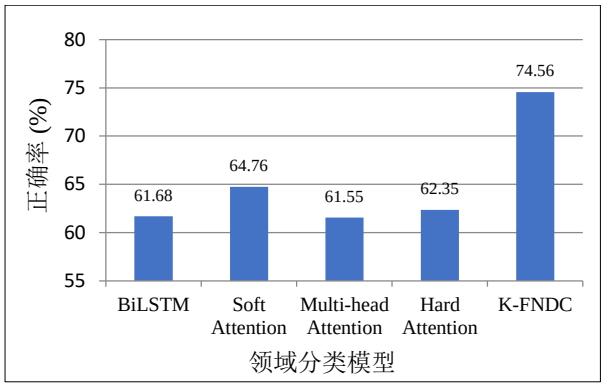

图 3　不同模型在知识敏感型领域性能对比

4　结束语

本文提出了结合特殊领域实体识别的远监督话语领域分类模型。该模型利用来自外部的领域知识，通过细粒度的词级融合机制来丰富话语的表达。并结合特殊领域实体识别和本地补充性知识库，丰富基于远监督的知识表达。实验结果表明，相比于研究进展神经分类模型，本文提出的方法在 SMP-ECDT 基准语料库上取得了显著的效果，特别是在知识敏感型领域。在未来的工作中，我们将研究如何有效地降低远监督带来的标签噪声，即从一个有噪声的候选领域标签集中为实体指称项识别出正确的领域标签，进一步提高话语领域分类的性能。此外，将本文的方法应用到更多预训练方法（如 BERT、ELMO 等）得到的词（或者字）向量上也是有价值的工作。

参考文献

[1] 俞凯，陈露，陈博等．任务型人机对话系统中的认知技术——概念、进展及其未来 [J]．计算机学报，2015, 38 (12):2333-2348.

[2] G. Tur and R. De Mori. 2011. Spoken language understanding: Systems for extracting semantic information from speech [M]. John Wiley & Sons, Inc.

[3] G. Tür, L. Deng, D. Hakkani-Tür, et al. Towards deeper understanding: Deep convex networks for semantic utterance classification [C]. In Proceedings of the 37th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012), pp. 5045–5048.

[4] P. Y. Xu and R. Sarikaya. Contextual domain classification in spoken language understanding systems using recurrent neural network [C]. In Proceedings of the 39th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014), pp. 136–140.

[5] S. V. Ravuri and A. Stolcke. Recurrent Neural Network and LSTM Models for Lexical Utterance Classification. In Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015), pp. 135-139.

[6] K. S. Yao, B. L. Peng, Y. Zhang, et al. Spoken language understanding using long short-term memory neural networks [C]. In Proceedings of the 2014 IEEE Spoken Language Technology Workshop (SLT 2014), pp. 189–194.

[7] S. W. Lai, L. H. Xu, K. Liu, et al. Recurrent convolutional neural networks for text classification [C]. In Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI 2015), pp. 2267–2273.

[8] T. Mikolov, I. Sutskever, K. Chen, et al. Distributed Representations of Words and Phrases and their Compositionality [C]. In Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS 2013), pp. 3111–3119.

[9] S. Takamatsu, I. Sato and H. Nakagawa. Reducing wrong labels in distant supervision for relation extraction [C]. In Proceedings of the f the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2012), pp. 721-729.

[10] X. Ling and D. S. Weld. Fine-grained entity recognition [C]. In Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI 2012), pp. 94-100.

[11] X. Ren, W. Q. He, M. Qu, et al. AFET：Automatic fine-grained entity typing by hierarchical partial-label embedding [C]. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), pp. 1369-1378.

[12] L. H. Chen, J. Q. Liang, C. H. Xie, et al. Short text entity linking with fine-grained topics [C]. In Proceedings of

the 27th ACM International Conference on Information and Knowledge Management (CIKM 2018), pp. 457–466.
[13] J. D. Chen, A. Wang, J. J. Chen, et al. CN-Probase: A Data-driven Approach for Large-scale Chinese Taxonomy Construction [C]. In Proceedings of the 35th IEEE International Conference on Data Engineering (ICDE 2019), pp. 1706-1709.
[14] Y. Zhang and J. Yang. Chinese NER using lattice LSTM [C]. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), pp. 1554-1564.
[15] J. Wang, Z. Y. Wang, D.W. Zhang, et al. Combining knowledge with deep convolutional neural networks for short text classification [C]. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2018), pp. 2915-2921.
[16] Y. Deng, Y. Shen, M. Yang, et al. Knowledge as a bridge: improving cross-domain answer selection with external knowledge [C]. In Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018), pp. 3295-3305.
[17] M.G. Chen, X.M. Jin, and D. Shen. Short text classification improved by learning multigranularity topics [C]. In Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011), pp. 1776-1781.
[18] J. Silva, L. Coheur, A. C. Mendes, et al. From symbolic to sub-symbolic information in question classification [J]. 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] X. H. Phan, L. M. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections [C]. In Proceedings of the 17th International World Wide Web Conference (WWW 2008), pp. 91-100.
[21] P. Haffner, G. Tur, and J. H. Wright. Optimizing SVMs for complex call classification [C]. In Proceedings of the 28th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), pp. 632-635.
[22] C. Chelba, M. Mahajan, and A. Acero. Speech utterance classification [C]. In Proceedings of the 28th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), pp. 280-283.
[23] X. Yang, A. Loukina, and K. Evanini. Machine learning approaches to improving pronunciation error detection on an imbalanced corpus [C]. In Proceedings of the 4th IEEE Workshop on Spoken Language Technology (SLT 2014), pp. 300-305.
[24] R. Sarikaya, G. E. Hinton, and B. Ramabhadran. Deep belief nets for natural language call-routing [C]. In Proceedings of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011,), pp. 5680–5683.
[25] P. Y. Xu and R. Sarikaya. Convolutional neural network based triangular CRF for joint intent detection and slot filling [C]. In Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2013), pp. 78–83.
[26] Y. Kim. Convolutional Neural Networks for Sentence Classification [C]. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), pp. 1746-1751.
[27] S. Ravuri and A. Stolcke. A comparative study of recurrent neural network models for lexical domain classification [C]. In Proceedings of the 41th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2016), pp. 6075-6079.
[28] J. P. Cheng, L. Dong, and M. Lapata. Long short-term memory-networks for machine reading [C]. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), pp. 551–561.
[29] N. T. Vu, P. Gupta, H. Adel, et al. Bi-directional recurrent neural network with ranking loss for spoken language understanding [C]. In Proceedings of the 41th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), pp. 6060–6064.
[30] 柯子烜，黄沛杰，曾真. 基于优化“未定义”类话语检测的话语领域分类 [J]. 中文信息学报，2018，32(4): 105-113.
[31] B. Liu and I. Lane. Attention-based recurrent neural network models for joint intent detection and slot filling [C]. In Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016), pp. 685-689.
[32] Y. Kim, D. Kim, A. Kumar. Efficient large-scale neural domain classification with personalized attention [C]. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), pp. 2214–2224.
[33] Z. C. Yang, D. Y. Yang, C. Dyer, et al. Hierarchical attention networks for document classification [C]. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016), pp. 1480-1489.
[34] L. P. Heck, D. Hakkani-Tür and G. Tür. Leveraging knowledge graphs for web-scale unsupervised semantic parsing [C]. In Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013), pp. 1594-1598.
[35] Y. N. Chen, D. Hakanni-Tur, G. Tür, et al. Syntax or semantics? knowledge-guided joint semantic frame parsing [C]. In Proceedings of 2016 IEEE Spoken Language Technology Workshop (SLT 2016), pp. 348-355.
[36] C. Shi, S. J. Liu, S. Ren, et al. Knowledge-based semantic embedding for machine translation [C]. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), pp. 2245–2254.
[37] A. Moro, A. Raganato and R. Navigli. Entity linking meets word sense disambiguation: a unified approach [J]. Transactions of the Association for Computational Linguistics, 2014 (2): 231–244.
[38] J. D. Chen, Y. Z. Hu, J.P. Liu, et al. Deep short text classification with knowledge powered attention[C]. In Proceedings of the 33th AAAI Conference on Artificial Intelligence (AAAI 2019), pp. 6252-6259.
[39] M. Schuster and K. K Paliwal. Bidirectional recurrent neural networks [J]. IEEE Transactions on Signal Processing, 1997(45): 2673–2681.
[40] W. N. Zhang, Z. G. Chen, W. X. Che, et al. The first evaluation of Chinese human-computer dialogue

technology. Computing Research Repository, arXiv:1709.10217. Version 1.
[41] N. Srivastava, G. E. Hinton, A. Krizhevsky, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014(15): 1929–1958.
[42] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need [C]. In Proceedings of the 41th Annual Conference on Neural Information Processing Systems (NIPS 2017), pp. 6000-6010.
[43] S. Shankar, S. Garg, and S. Sarawagi, Surprisingly easy hard-attention for sequence to sequence learning [C], In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), pp. 640–645.

何宇虹（1994—），硕士研究生，主要研究领域为自然语言处理、知识图谱。
Email: hyhong @ stu.scau.edu.cn

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。
Email: pjhuang@scau.edu.cn

杜泽峰（1997—），本科，主要研究领域为自然语言处理。
Email: seeledu @stu.scau.edu.cn