计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191200088

基于 BERT 和 BiLSTM 的语义槽填充

张玉帅　赵　欢　李　博
湖南大学信息科学与工程学院　长沙 410082
(zhangyushuai@hnu.edu.cn)

摘　要　语义槽填充是对话系统中一项非常重要的任务,旨在为输入句子的每个单词标注正确的标签,其性能的好坏极大地影响着后续的对话管理模块。目前,使用深度学习方法解决该任务时,一般利用随机词向量或者预训练词向量作为模型的初始化词向量。但是,随机词向量存在不具备语义和语法信息的缺点;预训练词向量存在"一词一义"的缺点,无法为模型提供具备上下文依赖的词向量。针对该问题,提出了一种基于预训练模型 BERT 和长短期记忆网络的深度学习模型。该模型使用基于 Transformer 的双向编码表征模型(Bidirectional Encoder Representations from Transformers,BERT)产生具备上下文依赖的词向量,并将其作为双向长短期记忆网络(Bidirectional Long Short-Term Memory,BiLSTM)的输入,最后利用 Softmax 函数和条件随机场进行解码。将预训练模型 BERT 和 BiLSTM 网络作为整体进行训练,达到了提升语义槽填充任务性能的目的。在 MIT Restaurant Corpus,MIT Movie Corpus 和 MIT Movie trivial Corpus 3 个数据集上,所提模型得出了良好的结果,最大 F1 值分别为 78.74%,87.60%和 71.54%。实验结果表明,所提模型显著提升了语义槽填充任务的 F1 值。

关键词:语义槽填充;预训练模型;长短期记忆网络;上下文依赖;词向量
中图法分类号　TP391

Semantic Slot Filling Based on BERT and BiLSTM

ZHANG Yu-shuai,ZHAO Huan and LI Bo
College of Computer Science and Electronic Engineering,Hunan University,Changsha 410082,China

Abstract　Semantic slot filling is an important task in the dialogue system,which aims to label each word of the input sentence correctly. Slot filling performance has a marked impact on the following dialog management module. At present,random word vector or pretrained word vector is usually used as the initialization word vector of the deep learningmodel used to solveslot filling task. However,the random word vector has no semantic and grammatical information,and the pre-trained word vector only present one meaning. Both of them cannot provide context-dependent word vector for the model. We proposed an end-to-end neural network model based on pre-trained model BERTand Long Short-Term Memory network(LSTM). First,the pre-trained model (BERT) encoded the input sentence as context-dependentword embedding. After that,the word embedding served as input to subsequent Bidirectional Long Short-Term Memory network(BiLSTM). Andusing the Softmax function and conditional random field to decode prediction labels finally. The pre-trained model BERT and BiLSTM networks were trained as a wholein order to improve the performance of semantic slot filling task. The model achieves F1 scores of 78.74%,87.60% and 71.54% on three data sets(MIT Restaurant Corpus,MIT Movie Corpus and MIT Movie trivial Corpus) respectively. The experimental results show that our model significantly improves the F1 value of Semantic slot filling task.

Keywords　Slot filling,Pre-trained model,Long short-term memory network,Context-dependent,Word embedding

1　引言

智能对话系统已经成为学术界和工业界的研究热点,主要包括面向开放域的闲聊系统(如 Apple Siri 和微软小冰)和面向任务的对话系统(如 Google Assistant)。在面向任务的对话系统中,解析人类话语的过程被称作口语语言理解,主要任务包括意图识别和语义槽填充。意图识别是指分析用户话语所属的类别,属于分类任务。语义槽填充则是识别并标注特定领域关键词的属性值,属于序列标注任务。两者的最终目的是为对话处理模块提供意图和语义槽信息,以便对用户的需求提供更可靠的反馈,并帮助其完成特定领域的任务。比如,用户发出"查看长沙到北京的机票"命令,意图是预定机票,"长沙"的属性值是"出发地",而"北京"是"目的地",根据口语语言理解模块的处理结果,对话系统可以准确地回复长

到稿日期:2019-12-13　返修日期:2020-05-01　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划(2018YFC0831800)
This work was supported by the National Key R&D Project Program(2018YFC0831800).
通信作者:赵欢(hzhao@hnu.edu.cn)

沙到北京的机票信息。本文主要讨论语义槽填充任务。

语义槽填充的属性值标注方法采用 BIO(Beginning, Inside, Outside)标注策略。一个属性值(槽类型)会产生两个语义槽标签,以"B-槽类型"和"I-槽类型"的格式表示。表 1 列出了一个简单的示例,其中"B-"表示属于同一种槽类型短语的开始单词,"I-"表示属于同一种槽类型短语的后续单词,"O"标记无关词汇。

表 1 BIO 格式的数据标注示例

Table 1 Examples of data annotation in format BIO

单词	Where	is	cheap	pizza	in	New	York
标签	O	O	B-Price	B-Dish	O	B-Location	I-Location

随着深度学习技术的不断发展,长短期记忆网络(Long Short-Term Memory Network, LSTM)[1]在序列数据建模上表现出色,解决了一般的循环神经网络(Recurrent Neural Network, RNN)存在的长期依赖问题;基于 LSTM 的神经网络在序列标注任务方面得到了广泛应用,并取得了较佳的效果。为了有效地学习双向上下文信息,本文使用 BiLSTM 构建基本模型。

BERT[2]是在大规模无标注语料库上训练的深度双向语言表征的预训练模型,在多个自然语言处理任务上取得了最佳成绩。将不同自然语言处理任务下的数据输入 BERT, BERT 可以输出具有上下文依赖的单词特征表示,相比传统词向量表示(如 Word2Vec 和 GloVe 等)具有更丰富的语义信息,可以根据不同的自然语言处理任务产生更适合的特征表示,从而提高模型性能。

本文采用 BERT 产生的单词特征表示作为 BiLSTM 网络的输入词向量,构造基于 BERT 和 BiLSTM 的网络模型,建模语义槽填充任务,并在用于口语语言理解任务的 3 个数据集(MIT Restaurant Corpus, MIT Movie Corpus 和 MIT Movie trivial Corpus)上进行实验。

2 相关工作

语义槽填充任务的传统解决方法包括基于字典、基于规则和基于统计等[3]。基于统计的算法主要有最大熵马尔可夫模型(Maximum Entropy Markov Model, MEMM)[4]和条件随机场(Conditional Random Field, CRF)[5]。近年来,深度学习方法在口语语言理解领域得到了广泛应用。Mesnil 等[6]和 Xu 等[7]分别使用基于 RNN 和卷积神经网络(Convolutional Neural Network, CNN)的模型解决语义槽填充任务。Xu 等[8]提出了基于 BiLSTM 的 BiLSTM-CRF 模型。Yao 等[9]研究了单层 LSTM 在口语语言理解上的应用,并进一步研究了在 LSTM 输出层执行回归操作和使用多层 LSTM 解决语义槽填充的问题。Peng 等[10]指出 RNN 网络虽然拥有记忆长期依赖的特性,但由于简单的 RNN 网络存在梯度消失和梯度爆炸问题,缺乏有效记忆长句子的能力,并通过引入外部记忆来增强简单 RNN 的记忆有限问题。Vu 等[11]使用 CNN 构造了新奇的 bi-directional sequential CNN 模型,在不依赖外部知识的前提下,显著提高了语义槽填充 F1 值。Kurata 等[12]提出了 encoder-labeler LSTM 模型,使用一个编码 LSTM 网络将输入句子编码成固定维度的句向量,并将此句向量作为另一个 LSTM 网络的输入向量,以将整个句子信息融入网络。Liu 等[13]提出多领域对抗训练方法,学习多领域语义槽填充任务的共享特征和表示。实验表明,利用对抗训练方法得到的通用领域模型有助于提高具体领域的语义槽填充性能。Zhao 等[14]使用 Seq2Seq(Sequence to Sequence)的生成神经网络模型来解决语义槽填充任务,并且通过引入 pointer 网络较好地解决了口语语言的未登录词问题。口语语言理解中的输入句子存在很多口语化的单词,也称作开放性词汇,如 MIT Corpus 中的餐馆名字、电影名字等。为了解决这个问题,Kim 等[15]构造了基于注意力机制的双向 RNN 网络,并将特殊的噪声向量融入输入句子的词向量,使得网络可以学习开放域单词槽位的特征,提高了对开放域词汇的语义槽填充能力。Yoo 等[16]提出利用隐变量模型(如变分自编码)的生成能力构造生成式数据增强框架,以缓解数据缺乏问题,进一步提高了口语语言理解能力。Shin 等[17]使用变分自编码结构的深度生成模型生成带有槽值对和多样化上下文信息的训练数据,并通过数据增强的方法提高了语义槽填充性能。

3 模型结构

图 1 给出了基于 BERT 和 BiLSTM 的模型结构,其主要由 BERT 词嵌入层、BiLSTM 网络层和解码层构成。词嵌入层使用预训练模型 BERT 将输入序列转化成对应的词向量序列;BiLSTM 层利用前向网络和后向网络编码当前时刻的上下文信息;解码单元利用 Softmax 函数或者 CRF 输出每个单词所对应的标签。

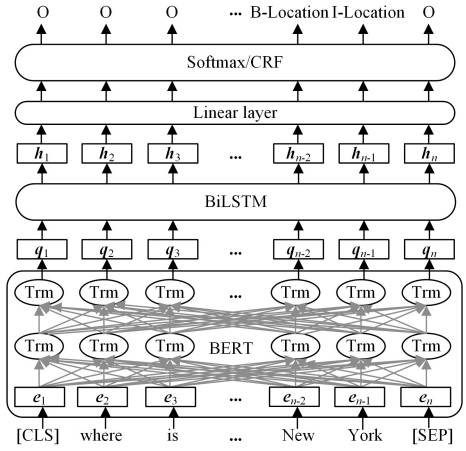


图 1 模型结构

Fig. 1 Model architecture

3.1 BERT 简介

BERT 是一个基于 Transformer 模型[18]的多层双向 Transformer 编码器,内部采用 Transformer 作为编码结构(ENCODER)。不同于 ELMo[19]利用双向 LSTM 作为编码单元,BERT 使用 Transformer,比传统的 RNN 具有更强的信息捕捉能力。为了学习更强的双向语言表征能力,BERT 模型在 BooksCorpus[20]和英文维基百科组合而成的数据集上进行了两个无监督的语言模型任务:掩码语言建模和下一个句子的预测。

3.2 BERT 产生词向量

BERT 的输入表示包含 3 部分，分别是词嵌入、位置嵌入和段嵌入。词嵌入是将句子经过 WordPiece Embedding[21] 处理之后的词表示。位置嵌入是 Transformer 模型中标识一个句子的单词顺序表示，最大值为 512。段嵌入是在成对句子的自然语言处理任务中区分前后句子的表示，“0”代表第一个句子，“1”代表第二个句子。语义槽填充任务只存在一个句子，因此本文实验中将段嵌入设置为“0”。

较为特殊的是，BERT 在每个句子的开始和结尾分别加入“[CLS]”和“[SEP]”标志，使得实际输入句子的长度增加了 2。其中，“[CLS]”位置对应的向量表示整个句子的句向量，一般用于句子分类任务；“[SEP]”是两个连续句子的分隔符。本文任务中没有涉及分类和成对句子任务，因此将这两个位置所对应的标签均设置为“O”。

令$[x_1, x_2, \cdots, x_n]$表示带有“[CLS]”和“[SEP]”的输入句子序列，n 为模型设置的输入句子的最大长度。BERT 的输入处理模块将每个x_i编码成一个向量$\boldsymbol{e}_i$：

$$\boldsymbol{e}_i = E_{token}(x_i) + E_{seg}(x_i) + E_{pos}(x_i) \tag{1}$$

其中，$E_{token}(x_i)$表示词嵌入，$E_{seg}(x_i)$表示段嵌入，$E_{pos}(x_i)$表示位置嵌入。

将输入处理模块输出的$[\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n]$作为 BERT 内部 Transformer 编码网络的输入词表示。预训练模型 BERT 通过加载预训练完成的参数，推理出输入序列的词向量表示$[\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_n]$。

3.3 双向长短期记忆网络

BiLSTM 的基础构件 LSTM 是 RNN 的一种变体。RNN 的每一个细胞的隐藏向量$\boldsymbol{h}_t$是由输入向量$\boldsymbol{q}_t$和前一个细胞的隐藏向量$\boldsymbol{h}_{t-1}$共同决定的，具体的计算过程如式(2)所示。

$$\boldsymbol{h}_t = \phi(\boldsymbol{q}_t \cdot \boldsymbol{w}_{qh} + \boldsymbol{h}_{t-1} \cdot \boldsymbol{w}_{hh'} + \boldsymbol{b}_h) \tag{2}$$

其中，$\phi()$是非线性激活函数，一般是 tanh 函数或 sigmoid 函数；$\boldsymbol{w}$和$\boldsymbol{b}$是可训练的参数。简单的 RNN 网络在反向传播中存在梯度爆炸或者梯度消失的问题，从而限制了网络学习长期依赖的能力。为了解决长期依赖问题，基于门控 RNN(gated RNN)的 LSTM 被提出，其基本结构如图 2 所示。

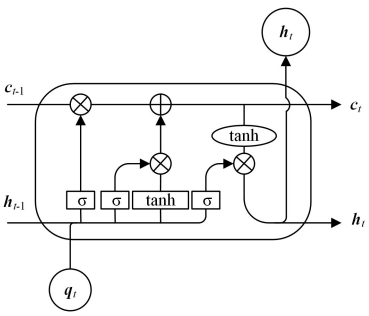


图 2 LSTM 细胞结构[22]

Fig. 2 LSTM cell architecture[22]

LSTM 引入了 3 个门控单元，分别是遗忘门$\boldsymbol{f}_t$、输入门$\boldsymbol{i}_t$和输出门$\boldsymbol{o}_t$，依次对应式(3)－式(5)。式(6)中，$\tilde{\boldsymbol{c}}_t$表示 t 时刻的输入信息。式(7)利用$\boldsymbol{f}_t$和$\boldsymbol{i}_t$控制细胞遗忘旧信息($\boldsymbol{c}_{t-1}$)和更新新信息($\tilde{\boldsymbol{c}}_t$)，从而得到当前时刻细胞向量$\boldsymbol{c}_t$；再利用$\boldsymbol{o}_t$输出细胞向量$\boldsymbol{c}_t$的重要信息，得到细胞隐藏向量$\boldsymbol{h}_t$(见式(8))。

$$\boldsymbol{f}_t = \sigma(\boldsymbol{q}_t \cdot \boldsymbol{w}^f_{qh} + \boldsymbol{h}_{t-1} \cdot \boldsymbol{w}^f_{hh'} + \boldsymbol{b}^f_h) \tag{3}$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{q}_t \cdot \boldsymbol{w}^i_{qh} + \boldsymbol{h}_{t-1} \cdot \boldsymbol{w}^i_{hh'} + \boldsymbol{b}^i_h) \tag{4}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{q}_t \cdot \boldsymbol{w}^o_{qh} + \boldsymbol{h}_{t-1} \cdot \boldsymbol{w}^o_{hh'} + \boldsymbol{b}^o_h) \tag{5}$$

$$\tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{q}_t \cdot \boldsymbol{w}^c_{qh} + \boldsymbol{h}_{t-1} \cdot \boldsymbol{w}^c_{hh'} + \boldsymbol{b}^c_h) \tag{6}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t \tag{7}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t) \tag{8}$$

其中，$\boldsymbol{q}_t \in \mathbb{R}^d$是 BERT 词嵌入层的输出词向量，$\odot$是向量逐点相乘运算，$\sigma$ 表示 sigmoid 激活函数，$tanh$ 是双曲正切激活函数。

通常，LSTM 网络从前向后编码句子，只掌握了从前到后的上下文信息，没有掌握从后到前的上下文信息，因此将前向 LSTM 网络和后向 LSTM 网络组成 BiLSTM 网络来学习双向上下文信息。网络结构如图 3 所示。

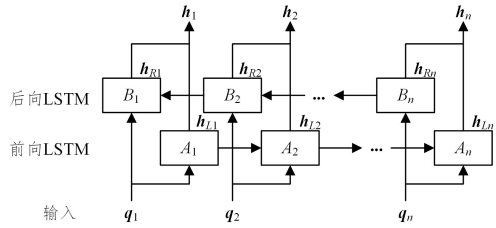


图 3 BiLSTM 网络结构

Fig. 3 BiLSTM network architecture

前向和后向 LSTM 网络的结构相同，但不共享参数。通过将前向网络输出隐藏向量矩阵$[\boldsymbol{h}_{L1}, \boldsymbol{h}_{L2}, \cdots, \boldsymbol{h}_{Ln}]$和后向网络隐藏向量矩阵$[\boldsymbol{h}_{R1}, \boldsymbol{h}_{R2}, \cdots, \boldsymbol{h}_{Rn}]$进行拼接，$\boldsymbol{h}_i = (\boldsymbol{h}_{Li}, \boldsymbol{h}_{Ri})$，得到 BiLSTM 网络层的隐藏向量矩阵 $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n]$，$\boldsymbol{H} \in \mathbb{R}^{n \times 2h}$。

3.4 解码单元

如图 1 所示，Linear Layer(线性层)将 BiLSTM 输出向量维度从 $2h$ 降至与语义槽标签总数相同的维度。假设 s 代表标签总数，第 i 个位置的线性层输出如式(9)所示。

$$\boldsymbol{m}_i = \sigma(\boldsymbol{W}_m \boldsymbol{h}_i + \boldsymbol{b}_m) \tag{9}$$

其中，$\boldsymbol{W}_m \in \mathbb{R}^{s \times 2h}$，$\boldsymbol{b}_m \in \mathbb{R}^s$。

(1)归一化指数函数

令$\boldsymbol{m}_i = [c_{i,1}, c_{i,2}, \cdots, c_{i,s}]$，使用归一化指数函数(Softmax 函数)将其转化成概率值表示。如式(10)所示，将$\boldsymbol{m}_i$中的每个分量归一化至(0,1)，此时的每一个分量近似表示成第 i 个单词属于某个标签的概率值。选取概率值最大的标签作为预测结果。

$$\hat{y}_{i,j} = \frac{\boldsymbol{e}^{c_{i,j}}}{\sum_{j=1}^{s} \boldsymbol{e}^{c_{i,j}}} \tag{10}$$

在模型训练过程中，利用最小化交叉熵损失函数策略学习最佳的模型参数。损失函数如式(11)所示，令$y_{i,j}$表示第 i 个单词属于第 j 个标签的真实概率，取值为 0 或 1，$\hat{y}_{i,j}$是通过式(10)得出的模型预测值。

$$loss = \sum_{i=1}^{n} \{ - \sum_{j=1}^{s} y_{i,j} log \hat{y}_{i,j} \} \tag{11}$$

(2)条件随机场

通常情况下，序列标注时的标签之间存在相互依赖关系，

使用条件随机场 CRF 优化预测标签序列是一种广泛的做法。Zhou 等[23]通过在 BiLSTM 的基础上添加 CRF 层，提高了语义角色标注任务的性能。语义槽填充作为序列标注问题，存在标签相互依赖的情况，如标签“I-location”的下一个标签不会是“I-price”。相对于 Softmax 函数只考虑当前单词对标签的影响进行标签预测，CRF 将学习到标签结构模型，并选择一条概率最大的标签序列作为解码结果。

对于输入序列$\vec{x}=(x_1,x_2,\cdots,x_n)$，$n$ 是序列的单词个数，其所对应的某一种标签序列$\vec{y}=(y_1,y_2,\cdots,y_n)$的概率为：

$$score(\vec{x},\vec{y})=\sum_{i=0}^{n}T_{y_i,y_{i+1}}+\sum_{i=1}^{n}S_{i,y_i} \tag{12}$$

其中，$\boldsymbol{T}$ 是转移特征矩阵，$T_{y_i,y_{i+1}}$ 表示从标签 y_i 转移到标签 y_{i+1} 的概率；$\boldsymbol{S}$ 是状态特征矩阵，S_{i,y_i} 指第 i 个单词的标签是 y_i 的概率。求得最大的 $score(\vec{x},\vec{y})$，即可求得句子$\vec{x}$所对应的全局最优的标签序列。

4 实验验证

4.1 实验环境

本实验在操作系统为 Ubuntu 16.04.6 LTS 的服务器上进行，CPU 型号为 Intel(R) Core(TM) i9-9820X CPU @ 3.30GHz，机器内存大小为 125 GB；使用深度学习框架 Pytorch 在一块内存容量为 11GB、型号为 GeForce RTX 2080 Ti 的英伟达 GPU 上进行训练和预测。

4.2 实验数据

实验数据采用麻省理工学院(MIT)计算机科学和人工智能实验室收集制作的 3 个数据集，它们主要被用在口语语言理解研究领域，包括餐馆领域的 MIT Restaurant Corpus(restaurant)、电影领域的简单数据集 MIT Movie Corpus(movie)和复杂数据集 MIT Movie trivial Corpus(trivial)。相关数据集的统计信息如表 2 所列。

表 2 MIT 数据集的统计信息

Table 2 Statistics for the datasets from MIT

	restaurant	movie	trivial
训练集	6894	8798	7035
验证集	766	977	781
测试集	1521	2443	1953
词汇表	3745	6604	10823
槽类型(标签数)	8(17)	12(25)	12(25)

4.3 实验设计与结果分析

4.3.1 实验设计

为了验证使用预训练模型的有效性，本文构造了 4 个对比模型，分别为 Random-BiLSTM，Random-BiLSTM-CRF，GloVe-BiLSTM 和 GloVe -BiLSTM-CRF。将对比模型和本文提出的模型 BERT-BiLSTM 和 BERT-BiLSTM-CRF 分别在表 2 所列的 3 个数据集上进行实验，并使用精确率(precision)、召回率(recall) 和 F1 值作为评价模型的标准，其中 F1 值为综合精确率和召回率来考虑模型性能的评价指标。3 个数值的计算采用 conlleval.pl[1] 工具。

模型训练中的超参数设置如下：所有模型的 LSTM 隐藏向量维度设置为 256；batch 大小为 32；dropout 率设置为 0.5；对于 restaurant 和 movie 数据集，输入句子的最大长度 n 设置为 50；对于 trivial 数据集，设置 n 为 80；模型 Random-BiLSTM，Random-BiLSTM-CRF，GloVe-BiLSTM 和 GloVe -BiLSTM-CRF 的学习率设置为 0.003，模型 BERT-BiLSTM 和 BERT- BiLSTM-CRF 的学习率设置为 0.00005。

采用 Adam 优化器训练所有模型参数，并使用 early stopping 方法确定模型训练的停止时刻。具体实施过程如下，当在验证集上的 F1 值连续 5 次没有升高趋势时，停止训练。利用 F1 值最大的那轮训练参数得到的模型在测试集上进行测试，得出 F1 值。

(1) Random-BiLSTM 模型和 Random-BiLSTM-CRF 模型利用数据集本身构建单词表，然后根据该单词表产生随机词向量，并将其作为 BiLSTM 的输入，该词向量在模型训练中也同时被训练。两者的区别在于，前者使用 Softmax 作为解码单元，后者使用 CRF 作为解码单元。

(2)GloVe-BiLSTM 模型和 GloVe-BiLSTM-CRF 模型采用在维基百科语料库上预训练后的 GloVe 词向量(200 维)作为模型的输入。

(3)BERT-BiLSTM 模型和 BERT-BiLSTM-CRF 模型将 BERT 预训练模型和 BiLSTM 网络结合在一起，通过训练 BiLSTM 网络参数和微调预训练模型 BERT 参数的方式，提高语义槽填充的性能。其中，预训练模型选用英文数据集的基础版本 BERT(12 层编码网络，输出词向量维度为 768，1.1 亿参数量，对大小写不敏感)。

4.3.2 实验结果分析

表 3－表 5 列出了各个模型分别在 3 个数据集上的精确率、召回率和 F1 值。

表 3 restaurant 数据集上的实验结果

Table 3 Experimental results on restaurant dataset

(单位：%)

模型	精确率	召回率	F1 值
Random-BiLSTM	69.40	74.48	71.85
Random-BiLSTM-CRF	74.47	76.45	75.45
GloVe-BiLSTM	70.80	75.79	73.21
GloVe -BiLSTM-CRF	77.41	77.09	77.25
BERT-BiLSTM(本文)	75.64	79.24	77.40
BERT-BiLSTM-CRF(本文)	**77.42**	**80.10**	**78.74**

从表 3 的实验结果可以看出，在 restaurant 数据集上，BERT-BiLSTM-CRF 在精确率、召回率和 F1 值指标上取得了最大值，分别为 77.42%，80.10%和 78.74%，综合指标 F1 值相对于表 3 中的前 4 个对比模型分别提高了 6.89%，3.29%，5.53%和 1.49%。表 4 的实验结果显示，BERT-BiLSTM-CRF 模型取得了最大精确率 87.32%，BERT-BiLSTM 模型取得了最大的召回率 87.98%和最大的 F1 值 87.60%。综合指标 F1 值较对比模型分别提高了 6.98%，4.60%，4.96%和 3.00%。

[1] https://www.clips.uantwerpen.be/conll2000/chunking/

表 4 movie 数据集上的实验结果

Table 4 Experimental results on movie dataset

（单位：%）

模型	精确率	召回率	F1 值
Random-BiLSTM	78.61	82.73	80.62
Random-BiLSTM-CRF	82.08	83.93	83.00
GloVe-BiLSTM	80.73	84.64	82.64
GloVe -BiLSTM-CRF	84.14	85.07	84.60
BERT-BiLSTM(本文)	87.22	**87.98**	**87.60**
BERT-BiLSTM-CRF(本文)	**87.32**	87.41	87.36

表 5 trivial 数据集上的实验结果

Table 5 Experimental results on trivial dataset

（单位：%）

模型	精确率	召回率	F1 值
Random-BiLSTM	57.92	64.33	60.96
Random-BiLSTM-CRF	68.35	66.83	67.58
GloVe-BiLSTM	61.52	67.69	64.46
GloVe -BiLSTM-CRF	69.71	68.98	69.34
BERT-BiLSTM(本文)	67.28	71.32	69.24
BERT-BiLSTM-CRF(本文)	**69.95**	**73.20**	**71.54**

在 trivial 数据集上，BERT-BiLSTM-CRF 模型的 3 个指标均取得了最好结果，分别为 69.95%，73.2% 和 71.54%。相较于对比模型，F1 值分别超出 10.58%，3.96%，7.08% 和 2.20%。

综上，在 restaurant、movie 和 trivial 数据集上，所提模型在精确率、召回率和 F1 值上都有很大的提升，能够有效提升语义槽填充的性能。

在对比模型中，Random-BiLSTM 模型和 Random-BiLSTM-CRF 模型使用随机初始化词向量作为特征提取网络的输入，该词向量的训练只局限于当前数据集，不具备语义信息，也无法表征上下文信息，亦无法有效表征超出当前词表的单词，故模型的性能表现不优。在 GloVe-BiLSTM 模型和 GloVe-BiLSTM-CRF 模型中，GloVe 词向量是利用 GloVe 算法在大规模维基百科数据集上进行预训练，生成的词向量包含语义和语法信息，对于意义相近的词，词向量也比较“近”。预训练数据的词表足够大，克服了超纲词的限制。但是，GloVe 词向量中的每一个单词只对应一个词向量，无法根据具体任务语料中的上下文产生更适合的词向量。相反，在 BERT-BiLSTM 模型和 BERT-BiLSTM-CRF 模型中，预训练模型 BERT 的参数可以在训练过程中根据具体语料的输入进行微调，从而为每个单词产生具有上下文依赖的词特征表示，达到相同单词在不同语境中表示不同含义的目的，消除了“一词一义”的缺点；并且 BERT 的预训练语料足够大，可以更好地识别不同语境中的单词含义，从而提升相应自然语言处理任务的性能。

如表 6 所列，BERT-BiLSTM 模型和 BERT-BiLSTM-CRF 模型准确地预测出了“wafflehouse”的标签为餐厅名字，而其他对比模型未能准确识别。原因在于，在英文维基百科语料中，“wafflehouse” 是其收录的一个餐厅名字词条，在维基百科上训练的 BERT 可能学习到了这组单词代表餐厅名字的语义信息，从而在表 6 所列实例的上下文语境中，“wafflehouse”表示的是餐厅名字，而不是菜系名字。

表 6 实例展示

Table 6 Examples

输入	will wafflehouse accept a prepaid visa gift card
正确标签	O B-Restaurant_Name I-Restaurant_Name O O B-Amenity I-Amenity I-Amenity I-Amenity
Random-BiLSTM	O B-Cuisine I-Cuisine O O B-Amenity I-Amenity I-Amenity I-Amenity
Random-BiLSTM-CRF	O B-Cuisine I-Cuisine O O B-Amenity I-Amenity I-Amenity I-Amenity
Glove-BiLSTM	O B-Cuisine I- Restaurant_Name O O B-Amenity I-Amenity I-Amenity I-Amenity
Glove -BiLSTM-CRF	O B-Cuisine I-Cuisine O O B-Amenity I-Amenity I-Amenity I-Amenity
BERT-BiLSTM(本文)	O B-Restaurant_Name I-Restaurant_Name O O B-Amenity I-Amenity I-Amenity I-Amenity
BERT-BiLSTM-CRF(本文)	O B-Restaurant_Name I-Restaurant_Name O O B-Amenity I-Amenity I-Amenity I-Amenity

4.4 与现有其他工作的对比

表 7 展示了本文模型与文献[13]、文献[16]和文献[17]在 restaurant，movie 和 trivial 数据集上的 F1 值比较。

表 7 与现有模型的 F1 值对比

Table 7 Compare with the existing model(F1 score)

（单位：%）

模型	restaurant	movie	trivial
文献[13]模型	74.47	85.33	65.33
文献[16]模型	73.00	82.90	65.70
文献[17]模型	76.17	86.04	69.13
BERT-BiLSTM(本文)	77.40	**87.60**	69.24
BERT-BiLSTM-CRF(本文)	**78.74**	87.36	**71.54**

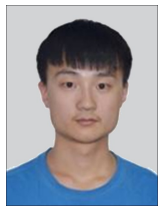
从表 7 可以看出，与多领域对抗方法[13]、数据增强方法[16]和话语生成方法[17]相比，BERT-BiLSTM 模型在 movie 数据集上取得了最佳的 F1 值，分别提高了 2.27%，4.7% 和 1.54%。BERT-BiLSTM-CRF 模型在 restaurant 数据集上取得了最佳的 F1 值，相比其他模型，分别提高了 4.27%，5.74% 和 2.57%；同时，在复杂电影数据集 trivial 上相比其他模型分别增长了 6.21%，5.84% 和 2.41%。实验结果表明，本文所提出的 BERT-BiLSTM 模型和 BERT-BiLSTM-CRF 模型有效地提升了语义槽填充性能。

结束语 语义槽填充任务旨在解析特定领域句子中的关键词，通常被当作序列标注任务。为了提高语义槽填充性能，本文以 BiLSTM 网络为基本模型，构建了 BERT-BiLSTM 模型和 BERT-BiLSTM-CRF 模型，使用 BERT 产生的具备上下文依赖的单词特征表示作为输入词向量，避免了传统词向量“一词一意”的缺点，提升了语义槽填充的准确率。在 restaurant，movie 和 trivial 3 个数据集上的实验结果表明，BERT-BiLSTM 模型和 BERT-BiLSTM-CRF 模型显著地提高了 F1 值，在一定程度上提升了语义槽填充性能。

参 考 文 献

[1] HOCHREITER S，SCHMIDHUBER J. Long short-term memo-

ry[J]. Neural Computation, 1997, 9(8): 1735-1780.
[2] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018.
[3] HOU L X, LI Y L, LI C C. Review of Research on Task-Oriented Spoken Language Understanding[J]. Computer Engineering and Applications, 2019, 55(11): 7-15.
[4] MCCALLUM A, FREITAG D, PEREIRA F C N. Maximum Entropy Markov Models for Information Extraction and Segmentation[C]// Proceedings of International Conference on Machine Learning. 2000: 591-598.
[5] RAYMOND C, RICCARDI G. Generative and Discriminative Algorithms for Spoken Language Understanding[C]// Proceedings of Conference of the International Speech Communication Association. 2008: 1605-1608.
[6] MESNIL G, DAUPHIN Y, YAO K, et al. Using Recurrent Neural Networks for Slot Filling in Spoken Language Understanding [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 23(3): 530-539.
[7] XU P, SARIKAYA R. Convolutional neural network based triangular CRF for joint intent detection and slot filling[C]// Proceedings of 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE, 2013: 78-83.
[8] XU Z X, CHE W X, LIU T. Slot filling based on Bi-LSTM-CRF [J]. Intelligent Computer and Applications, 2017, 7(6): 91-94.
[9] YAO K, PENG B, ZHANG Y, et al. Spoken language understanding using long short-term memory neural networks[C]// Proceedings of 2014 IEEE Spoken Language Technology Workshop(SLT). IEEE, 2014: 189-194.
[10] PENG B, YAO K. Recurrent Neural Networks with External Memory for Language Understanding[C]// Proceedings of Natural Language Processing and Chinese Computing. 2015: 25-35.
[11] VU N T. Sequential Convolutional Neural Networks for Slot Filling in Spoken Language Understanding[C]// Proceedings of 17th Annual Conference of the International Speech Communication Association(ISCA). 2016: 3250-3254.
[12] KURATA G, XIANG B, ZHOU B, et al. Leveraging Sentence-level Information with Encoder LSTM for Natural Language Understanding[J]. arXiv: 1601.01530, 2016.
[13] LIU B, LANE I. Multi-Domain Adversarial Learning for Slot Filling in Spoken Language Understanding [J]. arXiv: 1711.11310, 2017.
[14] ZHAO L, FENG Z. Improving slot filling in spoken language understanding with joint pointer and attention[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics(Volume 2: Short Papers). 2018: 426-431.
[15] KIM H Y, ROH Y H, KIM Y G. Data Augmentation by Data Noising for Open-vocabulary Slots in Spoken Language Understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop. 2019: 97-102.
[16] YOO K M, SHIN Y, LEE S. Data Augmentation for Spoken Language Understanding via Joint Variational Generation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 7402-7409.
[17] SHIN Y, YOO K M, LEE S G. Utterance Generation With Variational Auto-Encoder for Slot Filling in Spoken Language Understanding[J]. IEEE Signal Processing Letters, 2019, 26(3): 505-509.
[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of Advances in Neural Information Processing Systems. 2017: 5998-6008.
[19] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. arXiv: 1802.05365, 2018.
[20] ZHU Y, KIROS R, ZEMEL R, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 19-27.
[21] WU Y, SCHUSTER M, CHEN Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. arXiv: 1609.08144, 2016.
[22] JIN C, LI W H, JI C, et al. Bi-directional Long Short-term Memory Neural Networks for Chinese Word[J]. Journal of Chinese Information Processing, 2018, 32(2): 29-37.
[23] ZHOU J, XU W. End-to-end learning of semantic role labeling using recurrent neural networks[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015: 1127-1137.



ZHANG Yu-shuai, born in 1993, master, is a member of China Computer Federation. His main research interest is nature language processing.

ZHAO Huan, born in 1967, Ph.D, professor, is a member of China Computer Federation. Her main research interests include speech information processing, nature language processing and intelligent computing.