计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(7):75−83 [doi: 10.15888/j.cnki.csa.009157]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

改进 GL-GIN 的多意图识别和槽填充联合模型①

邓飞燕, 陈壹华, 陈禧琳, 李杰鸿

(华南师范大学 软件学院, 佛山 528225)
通信作者: 陈壹华, E-mail: YihuaChen1314@163.com

摘　要: 在当前自然语言处理多意图识别模型研究中, 存在建模方式均为从意图到插槽的单一方向的信息流建模, 忽视了插槽到意图的信息流交互建模研究, 意图识别任务易于混淆且错误捕获其他意图信息, 上下文语义特征提取质量不佳, 有待进一步提升等问题. 本文以当前先进的典型代表 GL-GIN 模型为基础, 进行优化改进, 探索了插槽到意图的交互建模方法, 运用槽到意图的单向注意力层, 计算插槽到意图的注意力得分, 纳入注意力机制, 利用插槽到意图的注意力得分作为连接权重, 使其可以传播和聚集与意图相关的插槽信息, 使意图重点关注与其相关的插槽信息, 从而实现多意图识别模型的双向信息流动; 同时, 引入 BERT 模型作为编码层, 以提升了语义特征提取质量. 实验表明, 该交互建模方法效果提升明显, 与原 GL-GIN 模型相比, 在两个公共数据集 (MixATIS 和 MixSNIPS) 上, 新模型的总准确率分别提高了 5.2% 和 9%.

关键词: GL-GIN; 多意图识别; 插槽填充; 联合模型

引用格式: 邓飞燕,陈壹华,陈禧琳,李杰鸿.改进 GL-GIN 的多意图识别和槽填充联合模型.计算机系统应用,2023,32(7):75–83. http://www.c-s-a.org.cn/1003-3254/9157.html

Multi-intent Detection and Slot Filling Joint Model of Improved GL-GIN

DENG Fei-Yan, CHEN Yi-Hua, CHEN Xi-Lin, LI Jie-Hong

(School of Software, South China Normal University, Foshan 528225, China)

Abstract: In the current research on multi-intention recognition models of natural language processing, information flow is only modeled from intention to slot, and the research on the interactive modeling of information flow from slot to intention is ignored. In addition, the task of intention recognition is easy to be confused, and other intention information is wrongly captured. The quality of contextual semantic feature extraction is poor and needs to be improved. In order to solve these problems, this study optimizes the current advanced typical GL-GIN (global-locally graph interaction network) model, explores the interactive modeling method from slot to intention, and uses the one-way attention layer from slot to intention. Furthermore, the study calculates the attention score from slot to intention, incorporates the attention mechanism, and uses the attention score from slot to intention as the connection weight. As a result, it can propagate and gather intention-related slot information and make the intention focus on the slot information that is relevant to it, so as to realize the bidirectional information flow of the multi-intention recognition model. At the same time, the BERT model is introduced as the coding layer to improve the quality of semantic feature extraction. Experiments show that the effect of this interactive modeling method is significantly improved. Compared with that of the original GL-GIN model, the overall accuracy of the new model on two public datasets (MixATIS and MixSNIPS) is increased by 5.2% and 9%, respectively.

Key words: global-locally graph interaction network (GL-GIN); multi-intent detection; slot filling; joint model

① 基金项目: 国家自然科学基金 (62076103)
收稿时间: 2022-12-22; 修改时间: 2023-02-03; 采用时间: 2023-02-08; csa 在线出版时间: 2023-04-17
CNKI 网络首发时间: 2023-04-18

在自然语言处理研究中, 口语语言理解 (spoken language understanding, SLU) 是对话系统的关键组成部分. 它包括意图识别 (intent detection, ID) 和插槽填充 (slot filling, SF) 两大核心子任务[1]. 意图识别旨在从给定语句中判断用户意图所属的最佳类别, 属于文本多标签分类任务; 而插槽填充则是指从给定的语句中识别提取插槽实体, 属于序列标记任务. 在意图识别和槽填充联合模型研究领域中, 根据意图识别数量进行分类, 当前研究方向可大致分为单意图 SLU 和多意图 SLU.

由于意图识别和插槽填充紧密相关, 虽然文献 [2–12] 中单意图 SLU 采用联合模型考虑两个任务之间的相关性, 也取得了显著的成功. 但由于多意图 SLU 指对话系统可以识别用户问题中含有的多个意图. 相比单意图 SLU, 多意图 SLU 在现实场景中更具有应用价值[13]. 因此关于多意图 SLU 技术的研究近年来受到越来越多的学者关注. 在此基础上, 2019 年, Gangadharaiah 等人[14] 首次尝试提出多任务框架来建立多意图识别和槽填充模型. 2020 年, Qin 等人[15] 设计了自适应交互框架来实现细粒度的多意图信息集成以填充插槽; 在此基础上, 2021 年, Qin 等人[13] 提出了全局-局部图交互网络 (global-locally graph interaction network, GL-GIN), 取得了目前最先进的性能效果. 成为本研究领域当前具有代表性的先进模型之一.

但是当前其依然存在以下问题.

(1) 仅考虑了从意图到插槽的单一方向的信息流建模, 忽视了插槽到意图的信息流交互建模.

(2) 多意图识别任务易于混淆且错误捕获其他意图的信息.

(3) 上下文语义特征提取质量不佳, 有待进一步提升.

基于此, 本文提出一种基于改进 GL-GIN 的多意图识别和槽填充联合模型 BIF-SI (bi-directional interaction framework with slot-intent), 另辟蹊径地探索了插槽到意图的交互方法, 实现传播和聚集与意图相关的插槽信息, 从而使意图重点关注与其相关的插槽信息, 避免错误混淆地捕获其他意图的信息, 以实现插槽到意图双向信息流交互建模, 将单向信息流动的 GL-GIN 模型改进为双向信息流动的交互模型 BIF-SI; 同时, 为进一步提高输入文本的语义特征提取质量, 引入 BERT 模型作为编码层.

1 相关工作

1.1 GL-GIN 多意图识别和槽位填充联合模型

GL-GIN 模型为 Qin 等人[13] 提出的非自回归模型, 即一种全局-局部图交互网络 (GL-GIN), 利用意图信息显式地指导槽填充任务, 实现当前最佳的性能.

如图 1 所示, GL-GIN 模型结构主要分为 4 个部分: 自注意力编码层, BiLSTM 层, 字符级别的意图解码层以及全局-局部图交互层. 其中, 以全局-局部图交互层最为复杂, 也是模型的核心部分, 全局-局部图交互层由两个主要部分组成: 一个是局部槽感知图交互网络, 用于建模跨槽依赖; 另一个是全局意图-插槽交互网络, 用于考虑意图和槽之间的交互, 其实质是利用图注意力网络 (graph attention network, GAT)[16] 的注意力层捕捉槽与意图之间的关系, 利用意图信息显式地指导槽填充任务, 实现了意图到槽的单向信息流动建模, 是当前多意图 SLU 研究中的先进代表模型, 其先进性能效对比如表 1 所示.

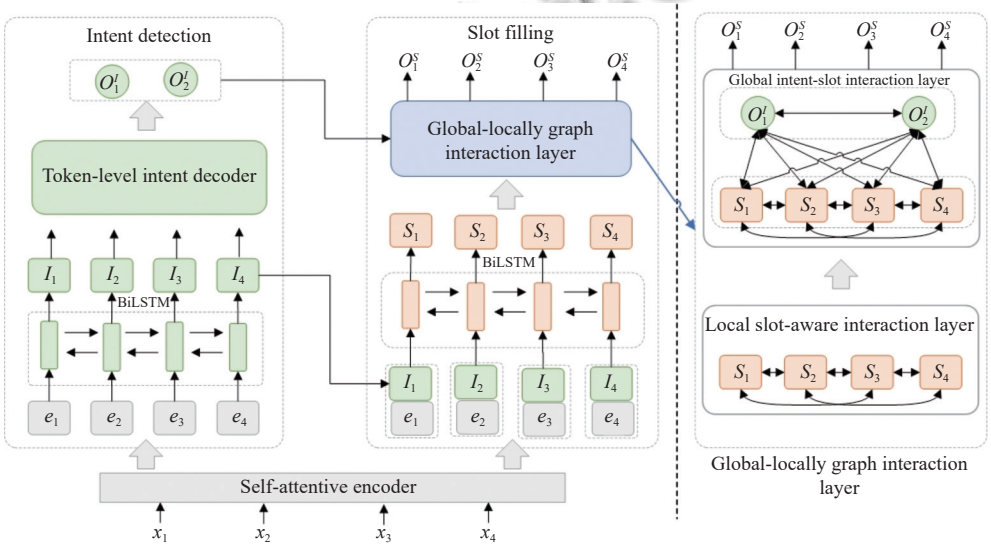

图 1 GL-GIN 模型架构图

表 1 当前代表性模型性能对比结果 (%)

模型	MixATIS			MixSNIPS		
	总准确率	插槽*F*1值	意图准确率	总准确率	插槽*F*1值	意图准确率
joint multiple ID-SF[14]	36.1	84.6	73.4	62.9	90.6	95.1
AGIF[15]	40.8	86.7	74.4	74.2	94.2	95.1
GL-GIN[13]	**43.5**	**88.3**	**76.3**	**75.4**	**94.9**	**95.6**

然而, 模型依然存在前述缺陷, 正如图 2 所示, 大多数现有的方法都重点关注了意图到插槽的信息流交互建模, 而忽视了插槽到意图的信息流交互建模. 实际上, 两者既密切相关, 又相互促进, 插槽信息同样可促进意图识别任务. 以一个意图识别案例为例, 如图 3 案例所示, 绿色为正确意图, 而红色则为错误的. GL-GIN 错误预测为“atis_flight”. 经观察发现, “atis_flight”为“la guardia”的意图, 而并非属于“la”的意图. 造成该错误预测的主要原因为上下文信息被平等对待, 从而混淆地捕获其他意图的信息. 而“B-city_name”“B-airport_name”和“I-airport_name”插槽信息对意图有一定的指示意义. 因此若能充分利用插槽信息促进意图识别任务, 让意图重点关注相关的信息, 有望避免此类错误的发生.

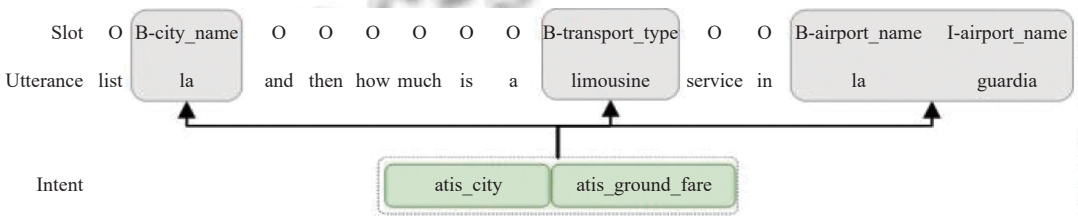

图 2 现有方法示意图

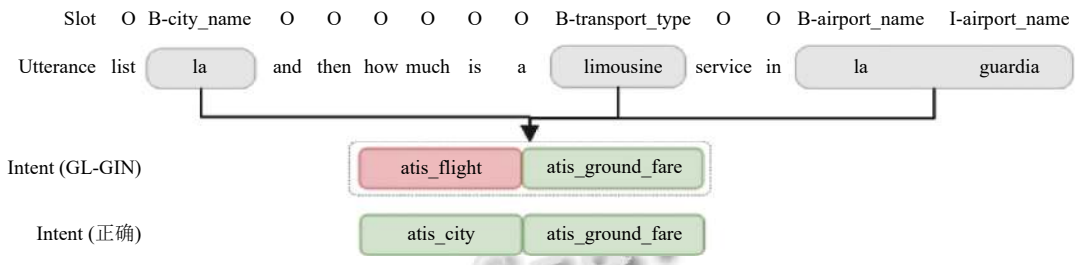

图 3 GL-GIN 意图识别案例示意图

2 改进 GL-GIN 模型

在 GL-GIN 模型中 , 如图 1 所示, 只有意图到插槽的信息流, 而没有插槽到意图的信息流. 而实际上, 这两个任务之间是可以相互促进的, 本文正是利用这个特点对 GL-GIN 模型进行改进, 探索提出双向信息流动的交互模型改进方案, 如图 4 所示.

针对多意图识别任务中, 上下文信息被平等对待, 受到其他无关信息的干扰, 导致混淆地捕获其他意图的信息, 从而影响意图识别准确率的问题, 在模型架构中增加 slot-to-intent 注意力层和 A-GCN 层, 运用槽到意图的单向注意力层, 计算插槽到意图的注意力得分, 使图卷积网络 (graph convolutional network, GCN) 模型纳入注意力机制, 改进为注意力图卷积网络 (attentional graph convolutional network, A-GCN), 利用插槽到意图的注意力得分作为连接权重, 使其能够传播和聚集与意图相关的插槽信息, 从而令意图重点关注与自身有关的插槽信息, 避免混淆地错误捕获其他无关的信息. 同时, 此改进方案成功实现了插槽到意图的信息流交互建模, 将单向信息流动的 GL-GIN 模型改进为双向信息流动的交互模型 BIF-SI.

此外, 针对 GL-GIN 模型采用自注意力编码层对输入文本进行上下文的语义特征提取, 特征提取质量尚存在一定的提升空间的可能性. 考虑到预训练 BERT 模型可以增强语义信息, 引入 BERT 模型作为编码层,

以进一步提高输入文本的语义特征提取质量.

综上所述, 本文对 GL-GIN 模型做了如下的优化改进, 如图 4 中黄色部分标识, 在原有模型架构的基础上, 引入以下 3 个组成部分.

1) BERT 编码层: 用以提高输入文本的语义特征提取质量.

2) slot-to-intent 注意力层: 用以计算插槽到意图的注意力得分, 作为 A-GCN 层的连接权重.

3) A-GCN 层: 用以传播和聚集与意图相关的插槽信息.

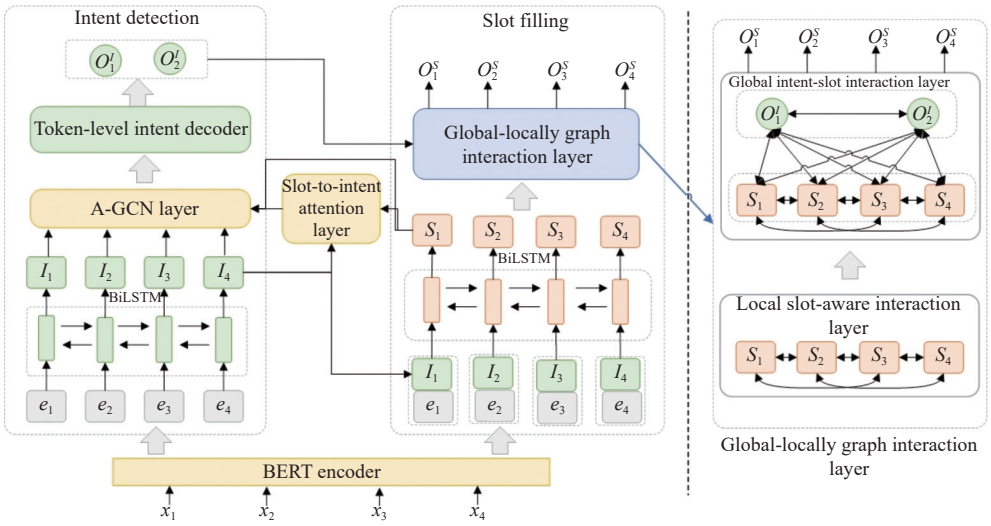

图 4 BIF-SI 模型架构图

图 4 为新构建的交互建模方案模型 (BIF-SI 模型) 架构图, 其结构主要分为 6 个部分: BERT 编码层, BiLSTM 层, slot-to-intent 注意力层, A-GCN 层, token 级意图识别解码层以及全局-局部图交互层.

2.1 BERT 编码层

采用 BERT 模型作为编码层, 使用 Hugging Face 中的 bert-base-uncased 预训练模型[17] 获得输入文本序列的向量表示, 以提高语义特征提取质量. bert-base-uncased 包含 12 层 Transformer, 每层 Transformer 都包含 12 头自注意力机制. 定义输入语句序列为$x=(x_1,\cdots,x_n)$, 其中x_i表示句子中的第i个字 (token). 在句子的起始位置加入 [CLS] 标签, 在句子的结尾处加入 [SEP] 标签, 经过预训练模型 BERT 处理后得到输入句子x的初始向量, 其表示为$E=\{e_1,\cdots,e_n\}\in\mathbb{R}^{n\times 2d}$.

2.2 BiLSTM 层

采用双向 BiLSTM 读取编码层输入向量$\{e_1,\cdots,e_n\}$, 以生成输入序列向前和向后上下文的隐藏状态表示. 定义 BiLSTM 层输出的意图感知隐藏表示和插槽感知隐藏表示分别为$I=(I_1,\cdots,I_m)$, $1\leqslant i\leqslant m$和$S=(S_1,\cdots,S_n)$, $1\leqslant j\leqslant n$, 其中 m 为意图的总个数, n 为插槽的总个数. 计算公式如下:

$$I_t=\text{BiLSTM}(e_t,I_{t-1},I_{t+1}) \tag{1}$$

$$S_t=\text{BiLSTM}(I_t\|e_t,S_{t-1},S_{t+1}) \tag{2}$$

其中, I_t和S_t分别表示第 t 个词的意图表示和插槽表示, $\|$表示拼接操作.

2.3 slot-to-intent 注意力层

在双向注意力实体图卷积网络 (bi-directional attention entity graph convolutional network, BAG)[18] 中, 双向注意力负责生成图节点和查询语句之间的交互信息, 被应用在多跳推理问答中. 实际上, 它在插槽和意图之间的注意力交互表现也较佳. 与 BAG 不同, BIF-SI 仅构建插槽到意图的单向注意力, 计算步骤如下.

(1) 计算每个插槽与意图之间的相似度. 相似度矩阵$s^*\in\mathbb{R}^{n\times m}$计算公式如下:

$$s^*=avg_{-1}f_a(concat(S,I,S*I) \tag{3}$$

其中, $S\in\mathbb{R}^{n\times d}$, $I\in\mathbb{R}^{m\times d}$分别是 BiLSTM 层输出的插槽表示和意图表示, d是意图和插槽表示转换后的共同维度, f_a是线性变换, avg_{-1}表示最后一个维度中的平均操作, $*$是元素级乘法.

(2) 计算插槽到意图的注意力得分. 与 BAG 中的

图节点到查询语句的注意力不同, BIF-SI 引入了一个插槽到意图的注意力得分矩阵$\tilde{a}_{S2I}=\{a_{1,1},\cdots,a_{n,n}\}$, $\tilde{a}_{S2I}\in\mathbb{R}^{n\times d}$, 保留与意图相关的插槽信息, 其计算公式如下:

$$\tilde{a}_{S2I}=Softmax_{col}(s^{*})\cdot I \tag{4}$$

其中, $Softmax_{col}$表示跨列执行$Softmax$函数, 而·表示矩阵乘法.

2.4 A-GCN 层

图卷积网络 (GCN) 是一种能够编码图中信息而被广泛使用的网络结构. 在每个 GCN 层中, 尽管它能够与邻居节点进行信息的传播与聚合, 但其词之间的连接均被平等地对待[19]. 即 GCN 模型不能合理区分不同连接的重要性. 为解决此问题, 使 GCN 纳入注意力机制, 改进为 A-GCN, 让其只传播和聚合与意图相关的插槽信息. 此改进具有重要意义, 其具体计算流程如下所示.

将插槽与意图的注意力得分矩阵$\tilde{a}_{S2I}$输入到 A-GCN 层, $a_{i,j}$作为连接权重, 从而专注于传播和聚集与意图相关的插槽信息, 如图 5 所示. A-GCN 层的输出h_t计算公式如下:

$$h_t=ReLU\left(a_{i,j}\left(W*\left(I_i+S_j\right)+b\right)\right) \tag{5}$$

其中, W和 b 分别为权重、偏置, $ReLU$ 为 $ReLU$ 激活函数.

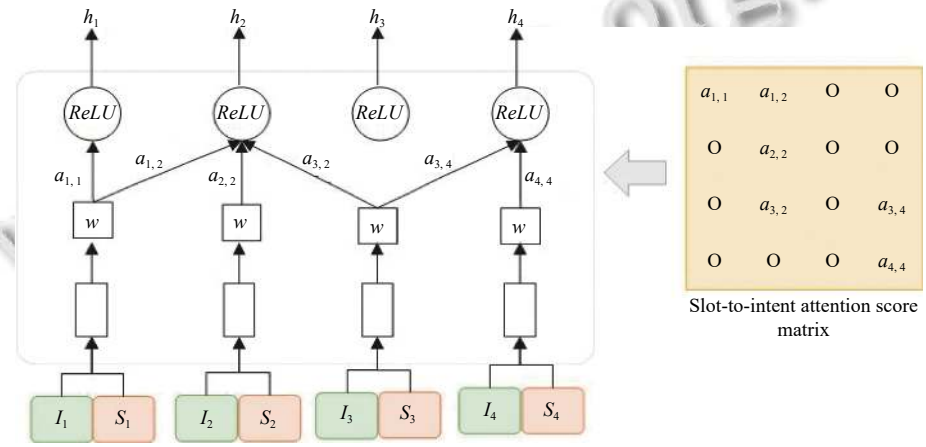

图 5 A-GCN 层

2.5 Token 级意图识别解码层

跟 GL-GIN[13] 类似, 在意图识别时, 采用了 token 级多标签多意图检测解码层, 通过对所有 token 投票得到语句意图结果. 具体过程如下所示.

首先, 利用 A-GCN 层的输出h_t进行意图检测. 第 t 个词的意图结果I_t^o计算公式如下:

$$I_t^o=\sigma(W_I(LeakyReLU(W_h h_t+b_h))+b_I) \tag{6}$$

其中, σ是 Sigmoid 激活函数, W_h和W_I是权重矩阵参数, b_h和b_I是偏置参数, $LeakyReLU$ 是 $LeakyReLU$ 激活函数.

然后, 当标签在 n 个 token 中得到超过一半的正面预测时, 则将此标签预测为语句意图, 语句意图结果o^I计算公式如下:

$$o^I=\left\{o_k^I\left|\left(\sum_{i=1}^{n}1[I_{(i,k)}^o>0.5]\right)>\frac{n}{2}\right.\right\} \tag{7}$$

其中, k 表示意图的数量, $I_{(i,k)}^o$表示的是 token i 对意图o_k^I的投票结果. 例如, 从 3 个 token 中获得关于 4 个意图的投票. 假设 3 个 token 对 4 个意图投票结果$I_{(i,k)}^o$分别为 I_1={0.9, 0.7, 0.8, 0.1}、I_2={0.9, 0.3, 0.7, 0.2}和 I_3={0.9, 0.2, 0.1, 0.1}; 如果$I_{(i,k)}^o$大于 0.5, 那么就累加 1, 计算最终的累加和, 即 4 个意图分别获得了{3, 1, 2, 0}的正面投票 (>0.5). 因此, 获得超过一半的选票 (>3/2)的标签为o_1^I和o_3^I, 最终预测的意图为 o^I={o_1^I, o_3^I}.

2.6 槽填充解码层

跟 GL-GIN[13] 一样, 在插槽填充时, 采用了全局-局部图交互网络, 这是一种非自回归方式, 实现了并行的槽填充解码.

2.6.1 全局-局部图交互层

全局局部图交互层主要由两部分构成: 一个是局部槽感知图交互网络, 用于建模时跨槽依赖; 另一个是全局意图-插槽图交互网络, 用于考虑意图和插槽之间的交互.

(1) 局部槽感知交互网络层. 每个字槽用一个顶点表示, 每个顶点都用相应的槽隐藏表示来进行初始化, 每个槽在一个窗口内部与其他槽建立连接, 由此构建出一个插槽和插槽之间交互的图. 槽之间的信息聚合

任务利用图注意力网络 (graph attention network, GAT)[16] 完成, 第 l 层的聚合过程可以定义为:

$$s_i^{l+1} = \sigma\left(\sum_{j\in N_i} \alpha_{ij} W_l s_j^l\right) \tag{8}$$

其中, N_i是一组表示连接插槽的顶点. 在叠加 L 层后, 得到上下文插槽感知的局部隐藏特征为$S^{L+1} = \{s_1^{L+1}, \cdots, s_n^{L+1}\}$.

(2) 全局意图-插槽交互网络. 每个序列槽连接所有预测的多个意图, 以实现并行输出槽序列. 用以下方法构造图$G = (V, E)$, V表示图的顶点, E 表示图的边.

① 顶点. 共有 $(n+m)$ 个节点, n 是序列长度, m 是意图解码器预测的意图标签数. 输入的槽的特征是上一步槽之间卷积得到的特征, 即$G^{[S,1]} = S^{L+1}$; 而意图节点的特征$G^{[I,1]} = \left\{\varphi^{\text{emb}}\left(o_1^I\right), \cdots, \varphi^{\text{emb}}\left(o_m^I\right)\right\}$由$o^I$变换得到, ϕ^{emb}表示可训练的嵌入矩阵; 插槽节点和意图节点的第 1 层状态向量为$G^1 = \left\{G^{[I,1]}, G^{[S,1]}\right\}$.

② 边. 在这个图网络中有 3 种类型的连接: 槽-意图, 槽-槽, 意图-意图.

③ 信息聚集. 全局 GAT 层的聚合过程可以表述为:

$$g_i^{[S,l+1]} = \sigma\left(\sum_{j\in G^S} \alpha_{ij} W_g g_j^{[S,l]} + \sum_{j\in G^I} \alpha_{ij} W_g g_j^{[I,l]}\right) \tag{9}$$

其中, G^S和G^I分别表示连接的槽和意图的顶点集.

2.6.2 插槽预测

经过 L 层的传播, 融合意图与槽的特征, 得到了用于槽预测的最终槽全局表示$G^{[S,L+1]}$.

$$y_t^S = Softmax\left(W_S g_t^{[S,L+1]}\right) \tag{10}$$

$$o_t^S = \text{argmax}\left(y_t^S\right) \tag{11}$$

其中, W_S是一个可训练参数, o_t^S是语句中第 t 个 token 的槽预测结果.

2.7 联合训练损失函数

受 Qin 等人[13] 的启发, 本文执行了一个联合训练损失函数来考虑多意图识别和槽填充任务. 通过联合优化来更新参数. 意图识别任务目标损失函数为:

$$CE(\hat{y}, y) = \hat{y}\log(y) + (1-\hat{y})\log(1-y) \tag{12}$$

$$\mathcal{L}_1 \triangleq -\sum_{i=1}^{n}\sum_{j=1}^{N_I} CE\left(\hat{y}_i^{(j,I)}, y_i^{(j,I)}\right) \tag{13}$$

同样, 槽填充任务目标损失函数为:

$$\mathcal{L}_2 \triangleq -\sum_{i=1}^{n}\sum_{j=1}^{N_S} CE\left(\hat{y}_i^{(j,S)}, y_i^{(j,S)}\right) \tag{14}$$

其中, N_I为单个意图标签的数量, N_S为插槽标签的数量.

最终目标损失函数公式为:

$$\mathcal{L} = \alpha\mathcal{L}_1 + \beta\mathcal{L}_2 \tag{15}$$

其中, α和β是超参数.

3 实验结果和分析

3.1 实验数据集

本文采用两个公开的多意图数据集进行性能对比实验和消融实验. (1) MixATIS[13,15], 包括 13 162 条语句用于训练, 756 条语句用于验证, 828 条语句用于测试. (2) MixSNIPS 数据集[13,15], 分别有 39 776、2 198、2 199 条语句用于训练、验证和测试. 本实验使用的是 Cleaned 版本的数据集, 删除了原始数据集中的重复句子, 该版本可在 https://github.com/LooperXX/AGIF 上找到.

3.2 实验设置

词嵌入的维度是 64, 意图和槽嵌入维度均为 128. LSTM 隐藏单元数为 256; 批处理大小为 16; 注意力机制的头数为 6; 图注意力网络的层数设置为 2; 学习率为 0.001. 为避免过拟合, 使用了 dropout 率为 0.4 的正则化方法. 使用 Adam[20] 优化模型中的参数. 对于所有实验, 本文选择在验证集上性能表现最好的模型, 然后利用测试集进行评估. 所有实验均在 NVIDIA Tesla V100 PCIe 16 GB 环境上进行.

同 Qin 等人[13] 的研究一样, 本文使用 $F1$ 值 ($F1$-score) 评估槽填充的性能, 使用准确率 (accuracy) 评估意图预测的性能, 使用总准确率 (overall accuracy) 评估意图和插槽都能被正确预测的语句所占比例.

3.3 实验对比模型

将改进模型与以下当前最佳的基线模型进行对比: (1) attention BiRNN[21]: 提出基于对齐的 RNN, 用于联合槽填充和意图识别; (2) slot-gated[2]: 提出一个槽门控机制的联合模型, 明确考虑槽填充和意图识别之间的相关性; (3) bi-model[22]: 提出一种多重交互的机制来不断增强意图识别和槽填充两个任务之间的联系; (4) SF-ID[5]: 提出了 SF-ID 网络以建立两个任务彼此之间的直接联系; (5) stack-propagation[7]: 采用 stack-propagation 框架明确结合意图识别以指导槽填充任务; (6) joint multiple ID-SF[14]: 提出一个含槽门控机制的多

任务框架, 用于多意图识别和插槽填充; (7) AGIF[15]: 提出了一种自适应交互网络来实现细粒度的多意图识别和槽填充; (8) GL-GIN[15]: 提出了一种全局-局部图交互网络, 实现了当前 SOTA 的性能.

3.4 对比实验结果与分析

为验证改进后模型性能的提升效果, 将所构建的模型与前述第 3.3 节所列举的现有代表性模型进行性能对比实验.

实验结果如表 2 所示, 有以下观察: (1) 在意图识别任务中, 去掉 BERT 后的 BIF-SI 模型在两个数据集上的意图准确率优于最佳基线模型 GL-GIN, 这表明本文提出的插槽到意图的建模方法能够成功应用于插槽信息指导意图识别任务, 从而提高意图识别任务的性能. (2) 更重要的是, 与 GL-GIN 相比, 改进后的模型 BIF-SI 在 MixATIS 和 MixSNIPS 数据集上, 其总准确率分别实现了 5.2% 和 9% 的提升. 原因是 BIF-SI 模型框架同时考虑了两个任务之间的交叉影响, 其中插槽信息可以用于促进意图识别任务; 此外, 引入 BERT 作为编码层可以进一步提高输入文本的语义特征提取质量, 从而提高 SLU 的性能.

表 2 不同模型性能对比实验结果 (%)

模型	MixATIS			MixSNIPS		
	总准确率	插槽*F*1值	意图准确率	总准确率	插槽*F*1值	意图准确率
attention BiRNN[21]	39.1	86.4	74.6	59.5	89.4	95.4
slot-gated[2]	35.5	87.7	63.9	55.4	87.9	94.6
bi-model[22]	34.4	83.9	70.3	63.4	90.7	95.6
SF-ID[5]	34.9	87.4	66.2	59.9	90.6	95.0
stack-propagation[7]	40.1	87.8	72.1	72.9	94.2	96.0
joint multiple ID-SF[14]	36.1	84.6	73.4	62.9	90.6	95.1
AGIF[15]	40.8	86.7	74.4	74.2	94.2	95.1
GL-GIN[13]	43.5	**88.3**	76.3	75.4	94.9	95.6
BIF-SI (w/o BERT)	43.8	87.5	76.9	75.8	93.9	**97.0**
BIF-SI (ours)	**48.7**	86.0	**79.6**	**84.4**	**96.5**	96.7

3.5 消融实验结果与分析

采用消融实验方法验证模型结构中改进方法的有效性. 在参数设置不变的条件下, 采用控制变量法实施消融实验, 即通过去掉该层结构, 观察其对模型效果降低程度的影响.

3.5.1 BERT 编码层的有效性

去掉 BERT encoder, 将其命名为“w/o BERT encoder”, 仍然采用原 GL-GIN 模型的 self-attentive encoder, 以验证 BERT encoder 层的有效性. 实验结果如表 3 所示. 在两个数据集中, 模型的总准确率分别下降了 4.9% 和 8.6%, 由此表明引入 BERT 模型作为编码层对 BIF-SI 模型效果提升具有巨大作用, 而这可归因于 BERT 预训练模型可以提供丰富的语义特征, 从而显著提高输入文本语义提取质量.

表 3 消融实验结果 (%)

模型	MixATIS			MixSNIPS		
	总准确率	插槽*F*1值	意图准确率	总准确率	插槽*F*1值	意图准确率
w/o BERT encoder	43.8	**87.5**	76.9	75.8	93.9	**97.0**
w/o slot-to-intent attention layer	47.8	86.0	77.7	83.8	96.5	95.7
w/o A-GCN layer	46.4	85.0	79.2	82.9	95.8	96.3
w/o slot-to-intent connection	45.3	84.9	77.9	82.9	96.3	96.7
BIF-SI (ours)	**48.7**	86.0	**79.6**	**84.4**	**96.5**	96.7

3.5.2 slot-to-intent 注意力层的有效性

去掉 slot-to-intent 注意力层, 将其命名为“w/o slot-to-intent attention layer”, 并直接将 BiLSTM 的输出提供给 A-GCN 层. 通过表 3 的实验结果可以清晰地观察到, w/o slot-to-intent attention layer 模型在 MixATIS 和 MixSNIPS 数据集上, 意图准确率分别下降了 1.9% 和 1%. 其性能的下降表明, 插槽到意图单向注意力层利用插槽信息显式指导意图识别任务起到了较好的作用, 原因是其可以通过计算每个插槽与意图之间的相似度和插槽到意图的注意力得分, 使得每个意图重点关注

与自身相关的插槽信息.

3.5.3 A-GCN 层的有效性

为了验证 A-GCN 层的有效性, 去掉 A-GCN 层, 利用插槽到意图单向注意力层的输出进行意图识别. 在表 3 中, 它被命名为“w/o A-GCN layer”. 可以观察到, 对于上述两个公开数据集, 该模型的意图准确率各自均下降了 0.4%, 这表明了 A-GCN 层将插槽到意图的注意力得分作为连接权重, 从而能够进一步传播和聚集与意图相关的插槽信息, 有利于提升 SLU 系统的语义性能.

3.5.4 双向连接 vs. 单向连接

为了验证模型双向连接建模的有效性, 只保留从意图到槽的信息流, 去掉槽到意图的信息流. 在表 3 中其被命名为“w/o slot-to-intent connection”. 根据实验结果可以发现到, 在 MixATIS 和 MixSNIPS 数据集上, 其总体精度分别降低了 3.4% 和 1.5%. 这说明从槽到意图方向的信息流交互建模可以带来更好的模型性能, 槽填充和意图识别之间的信息交互能够对这两个任务起到相互促进的作用. BIF-SI 在统一的框架下同时构建起两个任务之间的双向连接. 相比之下, 前人的工作只片面地考虑了单向信息流的交互作用.

3.6 定性分析实验

我们提供了一个案例对比研究, 以直观地理解插槽到意图的交互方法. 如图 6 所示, 绿色为正确意图, 而红色则为错误的. BIF-SI 正确地预测了“atis_city”, 经观察发现, “atis_city”为插槽“B-city_name”的意图, 其仅重点关注相关的插槽信息, 没有受到其他意图信息如“B-city_name”插槽信息的干扰. 而 GL-GIN 错误预测为“atis_flight”. GL-GIN 造成该错误预测的主要原因为上下文信息被平等对待, 从而混淆地捕获其他意图的信息. 相比之下, BIF-SI 的插槽到意图交互方法可以使意图重点关注与其相关的插槽信息, 从而利用插槽信息辅助意图识别任务, 有效避免其错误捕获其他无关的信息.

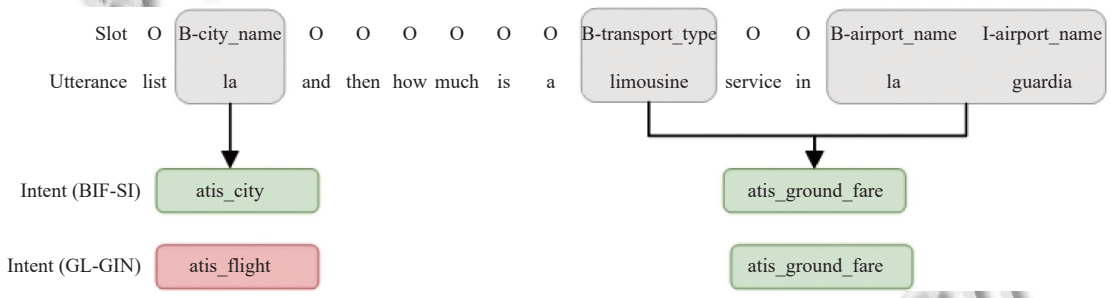

图 6 GL-GIN 和 BIF-SI 的案例对比研究

4 结论与展望

本文针对现有多意图识别联合模型忽视了插槽到意图的信息流交互建模, 多意图识别任务易于混淆地错误捕获其他意图的信息, 上下文语义特征提取质量有待进一步提升等问题, 提出一种基于改进 GL-GIN 的多意图识别和槽填充联合模型 BIF-SI. 具体来说, 探索了一种插槽到意图的交互建模方法, 应用插槽到意图的单向注意力层, 使得每个意图能重点关注与其相关的插槽信息; 同时利用 A-GCN 层进一步传播和聚集与意图相关的插槽信息, 从而使意图重点关注与自身相关的插槽信息, 避免混淆地错误捕获其他意图的信息. 此外, 引入 BERT 预训练模型作为编码层, 进一步提高输入文本语义特征提取质量; 实验结果表明, 改进后的模型在 MixATIS 和 MixSNIPS 数据集上的总准确率比先前的工作分别提高了 5.2% 和 9%.

当前模型性能还有一定提升空间, 并且本文只探索了槽位到意图的交互建模方法. 因此在未来的研究中, 将继续探索如何更加有效地对多意图和槽位双向交互进行建模, 进一步完善多意图 SLU 的工作.

参考文献

1 刘娇, 李艳玲, 林民. 人机对话系统中意图识别方法综述. 计算机工程与应用, 2019, 55(12): 1–7, 43. [doi: 10.3778/j.issn.1002-8331.1902-0129]

2 Goo CW, Gao G, Hsu YK, *et al*. Slot-gated modeling for joint slot filling and intent prediction. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: ACL, 2018. 753–757.

3 Li CL, Li L, Qi J. A self-attentive model with gate mechanism for spoken language understanding. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 3824–3833.

4 Liu YJ, Meng FD, Zhang JC, *et al*. CM-Net: A novel

collaborative memory network for spoken language understanding. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019. 1051–1060.

5 E HH, Niu PQ, Chen ZF, *et al*. A novel bi-directional interrelated model for joint intent detection and slot filling. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 5467–5471.

6 Zhang XD, Wang HF. A joint model of intent determination and slot filling for spoken language understanding. Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York: AAAI Press, 2016. 2993–2999.

7 Qin LB, Che WX, Li YM, *et al*. A stack-propagation framework with token-level intent detection for spoken language understanding. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019. 2078–2087.

8 Hakkani-Tür D, Tür G, Celikyilmaz A, *et al*. Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM. Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016). San Francisco, 2016. 715–719.

9 Xia CY, Zhang CW, Yan XH, *et al*. Zero-shot user intent detection via capsule neural networks. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 3090–3099.

10 Zhang CW, Li YL, Du N, *et al*. Joint slot filling and intent detection via capsule neural networks. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 5259–5267.

11 Wu D, Ding L, Lu F, *et al*. SlotRefine: A fast non-autoregressive model for joint intent detection and slot filling. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). ACL, 2020. 1932–1937.

12 Qin LB, Liu TL, Che WX, *et al*. A co-interactive transformer for joint slot filling and intent detection. Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021). Toronto: IEEE, 2021. 8193–8197.

13 Qin LB, Wei FX, Xie TB, *et al*. GL-GIN: Fast and accurate non-autoregressive model for joint multiple intent detection and slot filling. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. ACL, 2021. 178–188.

14 Gangadharaiah R, Narayanaswamy B. Joint multiple intent detection and slot labeling for goal-oriented dialog. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: ACL, 2019. 564–569.

15 Qin LB, Xu X, Che WX, *et al*. AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling. Proceedings of the 2020 Findings of the Association for Computational Linguistics: EMNLP 2020. ACL, 2020. 1807–1816.

16 Veličković P, Cucurull G, Casanova A, *et al*. Graph attention networks. arXiv:1710.10903, 2017.

17 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186.

18 Cao Y, Fang M, Tao DC. BAG: Bi-directional attention entity graph convolutional network for multi-hop reasoning question answering. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 357–362.

19 Tian YH, Chen GM, Song Y, *et al*. Dependency-driven relation extraction with attentive graph convolutional networks. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. ACL, 2021. 4458–4471.

20 Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014.

21 Liu B, Lane IR. Attention-based recurrent neural network models for joint intent detection and slot filling. Proceedings of the 17th Annual Conference of the International Speech Communication Association. San Francisco, 2016. 685–689.

22 Wang Y, Shen YL, Jin HX. A bi-model based RNN semantic frame parsing model for intent detection and slot filling. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: Association for Computational Linguistics, 2018. 309–314. [doi: 10.18653/v1/N18-2050]

(校对责编: 孙君艳)