软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(4):1885−1898 [doi: 10.13328/j.cnki.jos.006831]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种基于窗口机制的口语理解异构图网络*

张启辰, 王 帅, 李静梅

(哈尔滨工程大学 计算机科学与技术学院, 黑龙江 哈尔滨 150001)

通信作者: 张启辰, E-mail: zhangqichen@hrbeu.edu.cn

摘 要: 口语理解 (spoken language understanding, SLU) 是面向任务的对话系统的核心组成部分, 旨在提取用户查询的语义框架. 在对话系统中, 口语理解组件 (SLU) 负责识别用户的请求, 并创建总结用户需求的语义框架, SLU 通常包括两个子任务: 意图检测 (intent detection, ID) 和槽位填充 (slot filling, SF). 意图检测是一个语义话语分类问题, 在句子层面分析话语的语义; 槽位填充是一个序列标注任务, 在词级层面分析话语的语义. 由于意图和槽之间的密切相关性, 主流的工作采用联合模型来利用跨任务的共享知识. 但是 ID 和 SF 是两个具有强相关性的不同任务, 它们分别表征了话语的句级语义信息和词级信息, 这意味着两个任务的信息是异构的, 同时具有不同的粒度. 提出一种用于联合意图检测和槽位填充的异构交互结构, 采用自注意力和图注意力网络的联合形式充分地捕捉两个相关任务中异构信息的句级语义信息和词级信息之间的关系. 不同于普通的同构结构, 所提模型是一个包含不同类型节点和连接的异构图架构, 因为异构图涉及更全面的信息和丰富的语义, 同时可以更好地交互表征不同粒度节点之间的信息. 此外, 为了更好地适应槽标签的局部连续性, 利用窗口机制来准确地表示词级嵌入表示. 同时结合预训练模型 (BERT), 分析所提出模型应用预训练模型的效果. 所提模型在两个公共数据集上的实验结果表明, 所提模型在意图检测任务上准确率分别达到了 97.98% 和 99.11%, 在槽位填充任务上 *F*1 分数分别达到 96.10% 和 96.11%, 均优于目前主流的方法.

关键词: 对话系统; 口语理解; 异构图; 窗口机制; 意图检测; 槽位填充

中图法分类号: TP18

中文引用格式: 张启辰, 王帅, 李静梅. 一种基于窗口机制的口语理解异构图网络. 软件学报, 2024, 35(4): 1885–1898. http://www.jos.org.cn/1000-9825/6831.htm

英文引用格式: Zhang QC, Wang S, Li JM. Heterogeneous Graph Network with Window Mechanism for Spoken Language Understanding. Ruan Jian Xue Bao/Journal of Software, 2024, 35(4): 1885–1898 (in Chinese). http://www.jos.org.cn/1000-9825/6831.htm

Heterogeneous Graph Network with Window Mechanism for Spoken Language Understanding

ZHANG Qi-Chen, WANG Shuai, LI Jing-Mei

(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)

Abstract: Spoken language understanding (SLU), as a core component of task-oriented dialogue systems, aims to extract the semantic framework of user queries. In dialogue systems, the SLU component is responsible for identifying user requests and creating a semantic framework that summarizes user requests. SLU usually includes two subtasks: intent detection (ID) and slot filling (SF). ID is regarded as a semantic utterance classification problem that analyzes the semantics of utterance at the sentence level, while SF is viewed as a sequence labeling task that analyzes the semantics of utterance at the word level. Due to the close correlation between intentions and slots, mainstream works employ joint models to exploit shared knowledge across tasks. However, ID and SF are two different tasks with strong correlation, and they represent sentence-level semantic information and word-level information of utterances respectively, which means that the information of the two tasks is heterogeneous and has different granularities. This study proposes a heterogeneous interactive structure

* 收稿时间: 2022-05-09; 修改时间: 2022-08-08, 2022-09-20; 采用时间: 2022-11-03; jos 在线出版时间: 2023-06-14
CNKI 网络首发时间: 2023-06-15

for joint ID and SF, which adequately captures the relationship between sentence-level semantic information and word-level information in heterogeneous information for two correlative tasks by adopting self-attention and graph attention networks. Different from ordinary homogeneous structures, the proposed model is a heterogeneous graph architecture containing different types of nodes and links because a heterogeneous graph involves more comprehensive information and rich semantics and can better interactively represent the information between nodes with different granularities. In addition, this study utilizes a window mechanism to accurately represent word-level embedding to better accommodate the local continuity of slot labels. Meanwhile, the study uses a pre-trained model (BERT) to analyze the effect of the proposed model using BERT. The experimental results of the proposed model on two public datasets show that the model achieves an accuracy of 97.98% and 99.11% on the ID task and an *F*1 score of 96.10% and 96.11% on the SF task, which are superior to the current mainstream methods.

Key words: dialogue system; spoken language understanding (SLU); heterogeneous graph; window mechanism; intent detection; slot filling

1 引 言

面向任务的对话系统 (task-oriented dialogue system, TOD) 可以处理特定领域中的特定问题, 如智能聊天机器人、电影票预订等, 其中口语理解 (spoken language understanding, SLU) 是面向任务对话系统中的一个重要组件[1]. 面向任务的对话系统需要更严格的响应约束, 因为它的目标是根据用户信息进行精确的反馈. 在对话系统中, 口语理解组件负责识别用户的请求并创建一个简洁概括用户需求的语义框架. 该模块将原始用户消息转换为语义槽, 并对用户意图进行分类. 它通常涉及两个任务: 意图检测 (intent detection, ID) 和槽位填充 (slot filling, SF), 其分别用于识别用户意图和从自然语言表达中提取语义成分[2,3]. 意图检测被视为语义话语分类问题, 在句子级别分析话语的语义, 而槽位填充通常被视为在单词级别 (token-level) 工作的序列标记任务[4], 其性能将直接影响下游任务的决策.

例如, 图 1 为带有意图和槽注释 (BIO 格式) 的 SLU 话语示例, 槽位标签前缀“B-”表示标签是槽的开始, 标签前的前缀“I-”表示标签在槽内. “O”标记表示其他[5]. 话语如果检测到意图标签为“Flight”, 则单词“Kansas City”和“Newark”的槽位信息有可能被识别为“B-fromloc”“I-fromloc”和“B-toloc”. 但如果意图标签被识别为“Ground service”, 则上述槽位信息更有可能被识别为“B-City name”. 同时, 当槽位标签“B-fromloc”“I-fromloc”和“B-toloc”被填充时, 我们可以更准确地将意图信息识别为“Flight”而不是“Ground service”. 因此, 这两个任务之间存在很强的联系, 意图信息对槽位填充任务具有指导意义, 反之亦然.

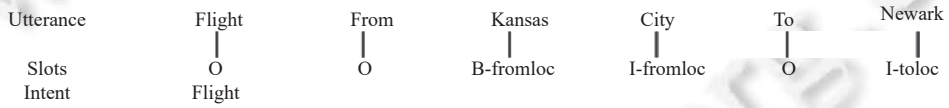


图 1 带有意图和槽注释 (BIO 格式) 的 SLU 话语示例

考虑到两个任务之间的显著相关性, 一些研究选择将意图检测和槽位填充任务结合到一个多任务学习框架中, 共同优化语义特征和共享潜在空间. 部分联合模型通过相互交互以促进意图检测和槽位填充任务的最终准确预测[4,6−9]. 这些模型具有显式控制两个任务的知识转移的优势, 可以帮助提高单词的可解释性, 同时有效分析 ID 和 SF 之间的影响[10]. 尽管这些模型取得了良好的效果, 但是这些模型使用同构结构, 没有考虑不同任务之间的特征差异. 因为意图检测是作用于整句话的句级语义分析任务, 而槽位填充是针对每个单词的词级任务, 它们所表示的特征是具有异构性的. 异构性是异构图的内在属性, 即拥有各种类型的节点和边, 不同类型的节点具有不同的特征, 其特征可能落入不同的特征空间中[11]. 异构图中的不同边可以提取不同的语义信息. 由于话语的每个词级 token 都是槽位填充任务的特定表示, 而意图检测是每个话语的分类任务, 它的表示是整体的. 同时, 还有一些模型没有注意到词意表达的局部性, 即在 SLU 中, 槽位不仅由关联项决定, 同时槽位标签“O”和“B-”“I-”具有局部连续性, 即“O”标签多数情况下呈局部出现, “I-”标签伴随着“B-”标签同时出现. 因此槽位信息会呈现出局部特征.

在本文中, 我们提出了一种异构结构框架来解决上述问题, 称为异构协同交互注意力网络 (heterogeneous co-interactive self-attention and graph attention network, HcoSG), 该模型是非自回归和协同交互的, 异构模型的核心采

用自注意力机制 (self-attention mechanism) 和图注意力网络 (graph attention network, GAT) 的联合形式来准确执行意图检测和槽位填充任务[12,13]. 图注意力网络 (GAT) 是一种新颖的卷积式图神经网络, 它是利用注意力机制来处理仅包含一种类型的节点或连接的非异构图. 事实上, 现实世界的图通常带有多种类型的节点和边, 也被广泛称为异构信息网络 (heterogeneous information network, HIN)[14]. 直观地说, 口语理解所构成的异构图中的 ID 和 SF 任务之间的关系可以有不同的语义, 分别代表语义级和词级. 同时, 由于意图检测是句子级别的分类任务, 而槽位填充任务是单词级别的分类任务, 意图检测相对于槽位填充是粗粒度的分类任务, 故这两个任务的表示信息具有不同的特征. 因此, 独立任务中的注意力机制应该与交互任务区分开来. 简而言之, 我们将这两个任务联合起来作为一个非自回归标签生成问题, 并且这两个任务相互迭代更新以摆脱不必要的时间依赖性. 与传统的自注意力机制和图注意力网络不同, 我们的模型是包含不同类型节点和连接的异构图结构, 可以使异构图涉及更全面的信息和丰富的语义. 基于学习到的异构结构的注意力值, 我们的模型可以获得邻居和多条边的最优权重组合且相互之间不共享, 从而使学习到的节点嵌入能够更好地捕捉异构图中复杂结构和语义信息的丰富性. 之后, 可以通过端到端的反向传播优化整个框架. 同时, 我们在槽位填充任务上采用窗口机制, 以更好地适应槽标签的局部连续性.

我们在两个公开数据集 ATIS[15]和 SNIPS[16]进行了实验, 针对两个数据集的实验结果都证明了我们框架及其各个组件的有效性, 实现了最先进的性能. 总而言之, 我们的贡献如下.

(1) 我们提出了一种异构结构框架, 对 ID 和 SF 任务进行联合建模, 以充分考虑不同任务类型的节点和连接所代表的不同语义信息. 据我们所知, 我们率先将异构图结构引入了 SLU 领域.

(2) 我们利用非自回归和窗口机制来准确表示话语的标记, 以更好地适应槽标签的局部连续性.

(3) 我们进行了广泛的实验来证明我们模型的有效性. 实验结果表明, 我们的模型在两个公共数据集上实现了最先进的性能. 同时, 我们采用了预训练模型 BERT, 以使得我们的模型效果进一步提升.

2 相关工作

在口语理解中, 意图检测一般被视为预测意图标签的语义分类问题, 而槽位填充主要被视为序列标注任务. SLU 模块将用户生成的自然语言消息转换为语义槽, 用于分类和意图检测. 早些年, 为了解决上述分类问题, 已经提出了一些方法, 如支持向量机 (support vector machine, SVM)[17]和条件随机场 (conditional random field, CRF)[18]. 最近, 基于深度学习的系统以其出色的性能引起了人们的关注.

对于意图检测任务模型, 深度凸网络[19,20]将前序神经网络的预测和当前话语结合起来, 作为当前网络的集成输入, 这种方法率先成功提高对话意图检测的准确率. 为了在序列处理中利用神经网络, 循环神经网络 (recurrent neural network, RNN) 和长短期记忆网络 (long short-term memory network, LSTM)[21,22]用作意图检测任务的话语编码器, 表明序列特征有利于意图检测任务. 最近, 预训练面向任务对话系统, 显著提高了意图检测子任务的预测准确性[23]. 该模型还表现出很强的稀缺数据学习能力, 可以有效缓解特定领域的数据不足问题.

槽位填充任务, 也称为语义标注任务, 是一个序列分类问题. 循环架构有利于序列标记任务, 因为它们可以沿着过去的时间步跟踪信息以充分利用序列信息. 常用的槽位填充神经网络方法包括条件随机场和循环神经网络, 基于 RNN 语言模型 (RNN-LMs)[24]可以用来检测序列标签, 而不是简单地预测单词, 其作者还对命名实体、句法特征和单词信息进行了研究. 有学者进一步研究了不同循环结构对槽位填充任务的影响, 发现所有 RNN 都优于 CRF 基线[25,26]. 与传统的序列标记方法不同, 文献 [27] 通过将其视为基于回合的跨度提取任务来解决槽位填充任务.

最近, 许多研究将意图检测和槽位填充任务结合到一个多任务学习框架中, 以共同优化共享潜在空间[28,29]. 一些方法考虑了从 ID 到 SF 的单一信息流, 因为意图信息可以为槽位填充提供句级语义特征. 门机制[30,31]的应用率先将意图信息应用于槽位填充任务. 堆栈传播框架[32]以执行单流操作令意图语义知识来指导槽位填充. Graph LSTM[7]方法利用时间步模拟意图和槽之间的语义相关性, 以达到两个任务信息的交互更新. 最近, Co-interactive transformer[6]和 CM-net[9]提出了交互意图检测和槽位填充任务的模型, 以充分利用两种信息的交互共享知识. TF[33]提出将语法知识编码到基于 Transformer 编码器的模型中, 用于意图检测和槽位填充, 语法监督可以帮助模型更好地学习语法模式. 这些联合模型可以准确地捕捉两个任务之间的共享知识, 从而整体提高两个任务的性能.

与上述工作相比, 我们的模型是一种异构图结构, 可以结合意图信息和槽位信息以充分考虑不同粒度的句级语义信息和词级信息之间的特征和差异. 同时, 我们利用了窗口机制来关注话语槽的局部连续性.

3 方 法

本节描述了我们提出的用于意图检测和槽位填充的联合模型. 该框架的架构如图 2 所示, 它由共享词级编码器、两级意图解码器、异构交互注意力层和意图感知槽填充解码器组成. 共享词级编码器的作用是形成词嵌入表示. 在两级意图解码器中, 首级阶段用作意图检测任务的预测, 末级阶段用来根据先验知识生成意图标签嵌入表示, 意图标签向量被作为异构交互注意力层的部分输入. 异构交互注意力层是该模型的核心部分, 用于整合两个任务的不同粒度信息和特征. 在意图感知槽填充解码器中, 意图预测信息被用来指导槽位特征信息的更新, 以更好适用于最终的槽值输出预测. 在本节中, 我们将详细介绍拟议框架的组成部分.

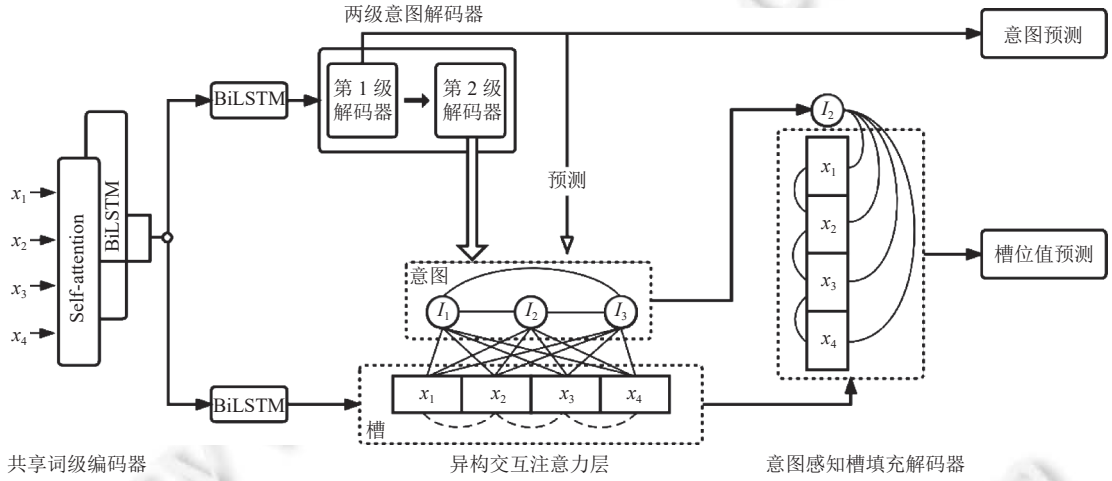


图 2 HcoSG 整体架构

3.1 共享词级编码器

在提出的框架中, 意图检测和槽位填充任务共享同一个词级编码器. 我们利用了自注意力机制和 BiLSTM (bi-directional long short-term memory) 编码的连接形式来获得可以整合词序时间特征和上下文信息的共享话语表示.

自注意力编码利用注意力机制和上下文感知功能分别实现局部和全局依赖. 在本文中, 我们利用自注意力机制来捕获每个 token 的上下文信息. 对于一个有 n 个单词的输入话语 $X=\{x_1,\ldots,x_n\}$, 每个 x_i 对应于一组 query、key 和 value 向量[12]. 自注意力机制通过将 x_i 的 query 向量与所有其余词嵌入表示的 key 向量一一相乘来计算每个 x_i 对 X 中所有其他词嵌入表示的注意力权重. 计算出的输出为 value 的加权和, 其中分配给每个 value 的权重由 query 与相应的 key 的 *Softmax* 函数计算. 形式化地, 自注意力输出 $A\in\mathbb{R}^{n\times d}$ (d 表示自注意力机制输出维度) 表示如下:

$$A=Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{1}$$

其中, Q, K 和 V 为序列 X 所进行线性变化后的矩阵, d_k 代表的是 key 的维度.

双向 LSTM (BiLSTM)[22]应用长期和短期记忆向量来编码顺序标记, 并使用门机制来控制信息流, 其广泛应用于序列标记问题. x_i 应用词嵌入函数 θ_{emb} 函数表示. BiLSTM 读取 x_i 以生成上下文感知隐藏状态序列:

$$H=\{h_1,\ldots,h_n\}\in\mathbb{R}^{n\times d},\ h_i=\left[\overrightarrow{h_i},\overleftarrow{h_i}\right]=\mathrm{BiLSTM}\left(h_{i-1},h_{i+1},\theta_{\mathrm{emb}}\left(x_i\right)\right) \tag{2}$$

我们将自注意力和 BiLSTM 的输出连接起来作为最终的编码表示:

$$E = A \,\|\, H = \{e_1, e_2, \ldots, e_n\} \tag{3}$$

其中, $E \in \mathbb{R}^{n\times 2d}$, 并且 $\|$ 表示连接操作.

3.2 两级意图解码器

在本节中, 我们执行两级意图解码器, 其中首级解码器用作最终的意图检测任务, 末级解码器用作生成意图标签的嵌入表示, 其可以更好地利用先验知识指导意图标签嵌入表示的生成. 准确对每个意图标签进行特征抽取使得槽位信息可以充分利用意图信息进行交互学习. 我们首先将在上一节中获得的上下文编码表示 E 输入到意图感知 BiLSTM 中, 以加强话语的任务特定表示:

$$h_i^I = \text{BiLSTM}\left(h_{i-1}^I, h_{i+1}^I, e_i\right),\ H^I = \left\{h_1^I, \ldots, h_n^I\right\} \in \mathbb{R}^{n\times d} \tag{4}$$

(1) 首级意图解码器

本阶段, 我们执行的是词级意图检测, 即在每个词的基础上预测当前话语的意图. h_i^I 被作为首级意图解码器的输入, 并且我们在 I_i 上执行了最大池化操作[34]以获得话语表示 $\tilde{I}_i$, 形式如下:

$$I_i = \text{LeakyReLU}\left(W_L h_i^I + b_L\right) \tag{5}$$

$$y_i^I = Softmax\left(W_I \tilde{I}_i + b_I\right) \tag{6}$$

其中, $I_i (I = \{I_1, \ldots, I_n\} \in \mathbb{R}^{n\times d})$ 表示第 i 个单词的意图预测向量, 其将用作末级意图解码器的输入, 以生成所有意图标签的嵌入表示. y_i^I 是每个单词的意图输出分布, 用于计算最终损失函数; W_I 和 W_L 是可训练矩阵, b_I 和 b_L 是偏差向量.

(2) 末级意图解码器

我们的目标是利用第 1 阶段生成的表示来创建意图标签嵌入以指导槽位填充任务. 单纯的随机初始化意图标签的嵌入表示再用于后续的预测任务是次优的, 因为其丢失了首级意图解码阶段针对意图预测的特定语义表示. 因此为了获得更丰富和鲁棒性的意图标签嵌入表征, 我们将第 1 阶段的词级意图输出进行一定变换操作以获得每个意图标签的特征信息, 其形式上:

$$\bar{I} = W_L' I + b_L' \tag{7}$$

$$\hat{I} = W_I'\left(\tanh\left(\bar{I}\right)^{\mathrm{T}}\right) + b_I' \tag{8}$$

其中, 我们使用 $\hat{I} \in \mathbb{R}^{|I_{\text{label}}|\times d_{\text{emb}}}$ 来表示意图嵌入, 并且 $|I_{\text{label}}|$ 表示意图标签的数目; $W_I' \in \mathbb{R}^{n\times d_{\text{emb}}}$ 和 $W_L' \in \mathbb{R}^{d\times|I_{\text{label}}|}$ 是可训练线性变换矩阵. 这意味着我们使用先验知识来表示每个意图标签. 由于意图标签的表示是通过第 1 阶段意图预测向量计算出来的, 它不是随机生成的, 故可以整合更丰富的意图标签编码信息. 因此, 第 2 阶段意图标签表示可以更准确地指导槽位填充任务.

3.3 异构交互注意力层

该部分是我们提出模型的核心. 由于意图检测是句子级的分类任务, 而槽位填充是词级的序列标注任务, 因此这两个高度相关的任务具有不同的粒度信息. 异构交互注意力层采用自注意力机制和图注意力网络的联合形式, 将意图表示和槽位词级嵌入表示输入到协同交互学习的统一框架中. 在这种结构中, 意图检测和槽位填充任务分别表征不同的语义, 代表句级语义特征和词级特征. 得益于这种注意力结构, 该框架可以同时考虑节点和路径的重要性. 同时, 我们利用窗口机制进行词级表示, 以更好地处理话语中的局部特征并降低模型复杂度. 下面, 我们依次描述带窗口机制的自注意力单元以及异构交互单元.

与第 3.2 节中的意图感知 BiLSTM 一致, 我们仍然使用 BiLSTM 来生成槽感知隐藏嵌入表示:

$$h_i^S = \text{BiLSTM}\left(h_{i-1}^S, h_{i+1}^S, e_i\right),\ H^S = \left\{h_1^S, \ldots, h_n^S\right\} \in \mathbb{R}^{n\times d} \tag{9}$$

(1) 带有窗口机制的自注意力单元

具有窗口机制的自注意力单元用于捕获每个单词的任务特定语义信息, 并充分利用槽信息的局部性. H^S 用作自注意力单元的输入, 以获得跨槽之间的依赖关系. 在窗口机制中, 窗口大小为定义为 δ, 表示可以关注当前隐藏节点的前序或后序相邻隐藏节点的数量, 词级节点基于自注意力单元进行更新:

$$Q_S, K_S, V_S = H^S W_q^{\mathrm{T}}, H^S W_k^{\mathrm{T}}, H^S W_v^{\mathrm{T}} \tag{10}$$

$$S = Softmax\left(f_{\mathrm{window}}\left(\frac{Q_S K_S^{\mathrm{T}}}{\sqrt{d_k}}\right)\right)V_S \tag{11}$$

$$f_{\mathrm{window}}(\cdot) = \begin{pmatrix} 1 & a_1 & a_2 & \cdots & a_{N-1} \\ a_1 & 1 & a_1 & \cdots & a_{N-2} \\ a_2 & a_1 & 1 & \cdots & a_{N-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{N-1} & a_{N-2} & a_{N-3} & \cdots & 1 \end{pmatrix}_{N\times N} \tag{12}$$

函数 $f_{\mathrm{window}}(\cdot)$ 表示一个用于计算注意力权重的 mask 矩阵, 其中 $a_i = 1, a_j = 0$ 当且仅当 $i \in \{1,\ldots,\delta\}$; $j \neq i$, 其表示一段序列中的一个词节点根据窗口大小 δ 对其相邻上下文节点产生的关注, 则针对窗口外的节点不产生关注, 即 mask 矩阵掩盖掉当前节点所不关注的节点权重, 再进行 *Softmax* 操作. 多头注意力允许模型共同关注来自不同位置的不同表示子空间的信息[12]. 同样, 我们采用多头注意力机制来达到更好的拟合效果. 更新后的槽位填充词级表示特征 $S = \{S_1,\ldots,S_n\} \in \mathbb{R}^{n\times d_{\mathrm{emb}}}$ 用于异构结构中部分迭代更新.

(2) 异构交互单元

图注意力网络是图神经网络的一种变体, 它被提出来学习节点与其邻居之间的重要性并融合邻居来进行节点分类. 在这里, 我们介绍了前面提到的词级自注意力机制, 它可以学习基于路径的邻居对异构结构中每个节点的重要性, 并将这些有意义的邻居的表示整合起来形成节点嵌入.

传统图注意力网络. 对于一个有 n 个节点的图结构, $\tilde{H} = \{\tilde{h}_1,\ldots,\tilde{h}_n\}$, $\tilde{h}_i \in \mathbb{R}^F$ 被定义为图注意力网络的初始输入始节点, 经过一层的信息交互更新后单层节点输出为 $\tilde{H}' = \{\tilde{h}'_1,\ldots,\tilde{h}'_n\}$, $\tilde{h}'_i \in \mathbb{R}^{F'}$. 在节点嵌入表示上操作的图注意力机制可以表示为:

$$\tilde{h}'_i = \mathop{\Big\|}_{k=1}^{K} \sigma\left(\sum_{j\in \mathrm{N}_i} \alpha_{ij}^k W_h^k \tilde{h}_j\right) \tag{13}$$

$$\alpha_{ij} = \frac{\exp\left(\mathrm{LeakyReLU}\left(a^{\mathrm{T}}\left[W_h\tilde{h}_i \,\|\, W_h\tilde{h}_j\right]\right)\right)}{\sum_{k\in N_i}\exp\left(\mathrm{LeakyReLU}\left(a^{\mathrm{T}}\left[W_h\tilde{h}_i \,\|\, W_h\tilde{h}_k\right]\right)\right)} \tag{14}$$

其中, σ 表示激活函数 (例如: Sigmoid 函数), $\|$ 表示连接操作; K 表示多头注意力头的数量; N_i 表示节点 i 的邻居节点域 (包括节点自身); $W_h \in \mathbb{R}^{F'\times F}$ 是可训练权重矩阵; $a \in \mathbb{R}^{2F'}$ 是可训练节点级注意力向量, 权重系数 α_{ij} 是非对称的, 这意味着两个节点之间对彼此的关注度不同. i 节点的基于路径的嵌入向量, 可以通过其邻居的投影特征聚合为 $\tilde{h}'_i$. 因为公式中分子拼接的序列不同, 以及它们有不同的相邻节点, 这个分母标准化项将会不同.

按照带窗口机制的自注意力单元的形式, 我们将图注意力网络和自注意力机制合并到异构结构中以整合不同的粒度信息, 并相互迭代更新. 如图 2 所示, 槽位填充任务的词级节点表示是利用 (1) 中提出的带窗口的自注意力机制和与之相连的意图节点通过自适应学习得到的, 意图节点嵌入表示也在同一个异构图中进行交互更新. 具体来说, 我们构造每个槽节点与意图标签表示 $\hat{I}$ 连接, $\hat{I}$ 来自第 2 级意图解码器. 这使框架能够对跨槽依赖进行建模, 从而缓解不协调的槽位问题. 我们通过以下方式构建图 $G = (V, E)$.

顶点 V: 当我们对槽和意图标记之间的交互信息进行建模时, 我们在异构图中有 $n+m$ 个节点, 其中 n 是话语的单词序列长度, $m = |I_{\mathrm{label}}|$ 是意图标签的数量. 意图标签嵌入表示和槽嵌入表示可以分别抽象地理解为句级信息和词级信息的特征.

边 E: 我们所提出的异构单元由 3 种类型的边组成.

(1) 由于意图检测和槽位填充任务高度相关, 我们连接意图节点 I_i 和 x_i 的槽节点以进行两个任务的信息交互. 具体来说, 每个槽节点连接所有意图标签节点以自适应地捕获相关意图信息.

(2) 我们互连意图标签节点并对每个意图标签之间的关系进行建模, 以了解彼此的语义级信息.

(3) 对于槽之间的连接, 我们应用了带有窗口机制的自注意力单元以进行槽节点之间的相互信息交互, 同时可以将其视为一种抽象边 (如图 2 中异构交互注意力层虚线所示).

通过这样, 我们在一个统一的框架中对意图和槽位信息进行显式建模. 意图标签节点的输入嵌入表示为 $\hat{I}=\{\hat{I}_1,\ldots,\hat{I}_m\}$ 是由第 2 级意图解码器生成的, 词级节点的输入嵌入表示为 $H^S=\{h_1^S,\ldots,h_n^S\}$. 交互信息过程形式化为:

$$S_i'=\mathop{\Vert}_{k=1}^{K}\left(\sigma\left(\sum_{j\in D_S}\alpha_{ij}^k W_g^k\hat{I}_j\right)+Softmax\left(f_{\text{window}}\left(\frac{h_i^S W_q^k(H^S W_k^k)^{\mathrm{T}}}{\sqrt{d}}\right)\right)\left(H^S W_v^k\right)\right) \tag{15}$$

$$I_i'=\mathop{\Vert}_{k=1}^{K}\sigma\left(\sum_{j\in D_I}\alpha_{ij}^k W_g^k\hat{I}_j+\sum_{j\in D_S}\alpha_{ij}^k W_g^k h_j^S\right) \tag{16}$$

其中, 公式 (15) 中的 α_{ij}^k 表示槽节点 h_i^S 对意图节点 $\hat{I}_j$ 的注意力权重, 与之类似的, 公式 (16) 中的 α_{ij}^k 表示意图节点 $\hat{I}_i$ 对意图节点 $\hat{I}_j$ 和槽位节点 h_j^S 的注意力权重, 以上 α_{ij}^k 均满足公式 (14) 的表现形式. 同时, 公式 (15) 中的自注意力算法为本节中提出的带有窗口机制的自注意力算法; $\sum_{j\in D_S}\alpha_{ij}^k W_g\hat{I}_j$ 和 $\sum_{j\in D_I}\alpha_{ij}^k W_g h_j^S$ 是用于合并语义级和词级交互信息的跨任务连接; 域 D_I 和 D_S 是表示意图标签节点和词级槽节点之间的连接边的顶点集合域, 其满足构造图中边 E 的 3 种连接形式. K 表示多头注意力头的数量. $S'=\{S_1',\ldots,S_n'\}\in\mathbb{R}^{n\times d'}$ 和 $I'=\{I_1',\ldots,I_m'\}\in\mathbb{R}^{m\times d'}$ 代表是迭代更新后的词级槽位和意图标签信息嵌入表示.

3.4 意图感知槽填充解码器

根据以上, 我们得到更新的节点信息, 其中每个槽节点已经包含了每个意图标签的特征. 为了实现更精确的槽位填充效果, 我们设置了本单元来利用两级意图解码器中每个单词的抽象意图表征 $\bar{I}$, 并经过一系列操作选取最有可能的意图标签, 并利用该意图信息来指导当前话语的最终槽位填充任务, 形式如下:

$$p_i^I=\max\left(\sum_{k=1}^{m}1\left[\sigma\left(\bar{I}_{(i,k)}\right)>0.5\right]\right) \tag{17}$$

$$O=\left\{\hat{I}_p,\hat{S}_1,\ldots,\hat{S}_n\right\}=\text{Transformer}\left(I_p\,\|\,S'\right) \tag{18}$$

其中, p_i^I 表示 $\bar{I}$ 经过变换后按维度相加并选择最大值作为学习到的意图嵌入特征表示的索引, 即 I_p 表示为 $I'=\{I_1',\ldots,I_m'\}$ 中索引为 p_i^I 的意图标签嵌入表示. Transformer$(\cdot)$ 表示自注意力机制对预测意图和槽节点嵌入拼接形式的最终解码 (形式上与带有窗口机制的自注意力算法一致, 窗口大小为 δ'), $\hat{S}=\{\hat{S}_1,\ldots,\hat{S}_n\}$ 是最终的槽位嵌入表示, 用作槽位填充任务预测. 我们应用标准条件随机场层[4]来解码槽标签:

$$O_s=W_S\hat{S}_i+b_S \tag{19}$$

$$y^S=\frac{\sum_{i=1}\exp f\left(y_{i-1},y_i,O_s\right)}{\sum_{y'}\sum_{i=1}\exp f\left(y_{i-1}',y_i',O_s\right)} \tag{20}$$

根据上述公式, W_S 为可训练矩阵, b_S 为偏执向量; $f(y_{i-1},y_i,O_s)$ 是计算从 y_{i-1} 到 y_i 的转换分数的函数, y^S 是预测的槽位填充标签序列.

3.5 联合训练

我们的模型经过训练以最小化意图检测和槽位填充的最终联合负对数似然目标函数. 其中意图检测和槽位填充目标损失函数如下:

$$L_1\triangleq-\sum_{j=1}^{n}\sum_{i=1}^{m}\hat{y}_j^{i,I}\log\left(y_j^{i,I}\right) \tag{21}$$

$$L_2\triangleq-\sum_{j=1}^{n}\sum_{i=1}^{T}\hat{y}_j^{i,S}\log\left(y_j^{i,S}\right) \tag{22}$$

其中, $\hat{y}_j^I$ 和 $\hat{y}_j^S$ 分别表示人工标注正确的意图和槽位标签; m 表示意图标签数量, T 表示槽标签数量. 最终联合损失函数如下:

$$L = \lambda L_1 + L_2 \tag{23}$$

其中, λ 为超参数.

4 实 验

4.1 数据集

为了全面评估我们提出的模型的性能, 我们在两个公共数据集 ATIS 和 SNIPS 上进行了实验.

ATIS[15]: 航空公司旅行信息系统 (airline travel information systems, ATIS) 数据集长期以来一直被用作口语理解的基准. 训练集包含 4 478 个话语, 验证集包含 500 个话语, 测试集包含 893 个话语, 共有 120 个不同的槽标签和 21 个不同的意图类型.

SNIPS[16]: 该数据集是从 SNIPS 个人语音助手中收集的, 具有 72 个槽位标签和 7 个意图类型. 训练集有 13 084 个话语, 验证集有 700 个话语, 测试集有 700 个话语.

4.2 实验设置及评估指标

在论文中, 对于 ATIS 和 SNIPS 数据集, 词嵌入的维数设置为 300. 两个数据集上的自注意力隐藏单元维数都是 1 024. 同时, 我们设置窗口大小 δ 为 2, δ' 为 1, 异构交互注意力层头数 K 设置为 8, 意图感知槽填充解码器中自注意力单元头数设置为 2. 超参数 λ 设置为 1. 我们框架上使用的 L_2 正则化为 1×10^{-6}, dropout 率设置为 0.2. 我们使用 Adam[35]去优化参数并设置训练次数为 100 次.

针对意图检测任务, 我们采用准确率 (accuracy) 来评估意图检测的预测性能:

$$\text{意图检测准确率(Intent Acc)}=\frac{\text{正确预测的样例个数}}{\text{样例总数}} \tag{24}$$

针对槽位填充任务, 我们采用 $F1$ 分数来评估槽位填充的预测性能:

$$\text{查准率}(P)=\frac{TP}{TP+FP} \tag{25}$$

$$\text{召回率}(R)=\frac{TP}{TP+FN} \tag{26}$$

$$\text{槽位填充}F1\text{分数}(F1)=\frac{2\times P\times R}{P+R} \tag{27}$$

其中, TP 为真正例个数, FP 为假正例个数, FN 为假反例个数.

同时, 我们采用总体精度 (overall accuracy) 针对句子级语义框架解析进行评估, 评估指标为:

$$\text{总体精度(overall accuracy)}=\frac{\text{意图和槽位都预测正确的样例个数}}{\text{样例总数}} \tag{28}$$

4.3 基线模型

为了全面评估我们所提出的模型 HcoSG, 我们将我们的模型与以下基线方法进行了比较.

Slot-Gated[30]提出了一种槽门机制, 可以专注于学习意图和槽注意力向量之间的关系, 以便通过全局优化获得更好的语义框架结果.

SF-ID Network[4]增强了双向关联连接, 为两个任务建立直接连接, 帮助它们相互促进.

CM-Net (collaborative memory network)[9]率先以协作的方式从记忆中捕获特定于槽位和特定于意图的特征, 然后使用这些丰富的特征来增强局部上下文表示, 在此基础上, 顺序信息流可以引出更特定的槽位和意图全局话语表示.

Stack-Propagation[32]执行 token 级别的意图检测, 以提高意图检测性能并进一步缓解错误传播.

Graph LSTM[7]提出用图长短时记忆网络来解决这个任务, 它首先将文本转换为图形, 然后利用消息传递机制

来学习节点表示.

Co-interactive Transformer[6]提出了协同交互模块, 通过在两个相关任务之间建立双向连接来考虑交叉影响, 其中槽和意图可以利用相应的互信息.

4.4 总体结果

与上述基线模型保持一致, 我们使用 *F*1 分数评估口语理解在槽位填充中的性能, 使用准确度评价意图检测的性能, 以及使用整体准确度进行句子级语义框架解析评估. 表 1 显示了我们提出的模型在 ATIS 和 SNIPS 数据集上的整体性能. 我们将我们提出的 HcoSG 的性能与所有基线方法进行了比较. 正如预期的那样, 我们的模型在两个公开数据集上的结果显示, 我们提出的模型优于以上的所有基线模型. 如表所示, 在 ATIS 数据集上, 与当下最优的联合建模模型 Co-interactive Transformer[6]相比, HcoSG 在槽位填充任务上的 *F*1 分数提高了 0.20%, 意图检测精度提高了 0.28%, 整体精度提高了 0.62%. 对于 SNIPS 数据集, 我们的框架在槽位填充任务上的 F1 分数超过了最先进的方法 0.21%, 在意图检测准确度上提高了 0.31%, 在整体准确度上提高 0.05%. 这证明了充分考虑不同任务类型的节点和连接所代表的不同语义信息的有效性. 我们将此归因于我们的模型充分利用了句级语义和词级异构信息结构, 可以更好地掌握意图和话语之间的关系, 使两个任务能够更充分地进行信息交互, 以提高模型的鲁棒性.

表 1 在 ATIS 和 SNIPS 数据集上 HcoSG 针对意图检测和槽位填充任务与基线方法的比较 (%)

Model	ATIS			SNIPS		
	Intent Acc	Slot *F*1	Overall Acc	Intent Acc	Slot *F*1	Overall Acc
Slot-Gated[30]	93.60	94.80	82.20	97.00	88.80	75.50
SF-ID Network[4]	97.09	95.80	86.90	97.29	92.23	80.43
CM-Net[9]	96.10	95.60	85.30	98.00	93.40	84.10
Stack-Propagation[32]	96.90	95.90	86.50	98.00	94.20	86.90
Graph LSTM[7]	97.20	95.91	87.57	98.29	95.30	89.71
Co-interactive Transformer[6]	97.70	95.90	87.40	98.80	95.90	90.30
HcoSG	**97.98**	**96.10**	**88.02**	**99.11**	**96.11**	**90.35**
HcoSG+BERT	98.10	97.61	90.21	99.55	97.45	93.38

4.5 单一组件消融实验与分析

我们推测我们提出的模型达到的改进效果是由于说话者的意图从槽节点级信息中获得了更多的知识, 而槽位填充任务也受益于会话的语义级意图信息, 这从本质上提高了对会话的上下文理解. 在本节中, 我们将从几个方向研究我们的模型. 我们首先进行了单一组件消融研究以检查我们框架中不同组件的效果和影响, 并分析自我注意窗口机制及窗口大小的影响.

(1) 两级意图解码器消融实验

我们认为, 两级意图解码器可以通过第 1 阶段对意图检测进行准确预测, 而第 2 阶段利用先验知识精确生成意图标签嵌入表示. 我们通过消除第 2 级意图解码器替换为使用可训练的嵌入矩阵来初始化生成意图标签嵌入表示作为异构交互层部分的输入来执行消融实验. 从表 2 中可以看出, 在 ATIS 和 SNIPS 数据集上槽 *F*1 分数分别下降了 0.53% 和 0.49%. 同时, 在两个数据集上的意图检测准确度和整体准确度也有所下降. 我们分析原因是模型没有准确生成意图标签表示, 使得意图表示没有表达更丰富的语义特征, 导致槽和意图之间的语义关联松散, 槽位填充任务的性能下降也会反作用于导致相应的意图检测任务准确性下降.

(2) 异构交互结构消融实验

此部分实验旨在验证异构结构的有效性. 我们将异构单元替换为同构结构, 即使用图注意力网络和自注意力机制分别进行意图信息和槽位信息的交互. 首先, 我们直接将意图标签嵌入表示与词级嵌入表示拼接起来, 再作为自注意力单元的输入, 以进行第 1 个消融实验. 在表 2 中将其称为仅使用自注意力机制. 我们可以观察到槽 *F*1 分数在两个公开数据集上分别下降了 0.10% 和 0.21%. 同时第 2 个消融实验, 我们同时将意图标签嵌入表示和词级

嵌入表示输入到一个图注意力网络中, 在表中我们称其为仅使用图注意力网络. 我们可以看到槽 $F1$ 分数也呈现下降趋势. 通过分析, 我们将异构交互结构的优势归功于自注意力机制可以准确捕捉词级节点之间的特征, 而图注意力网络可以将语义级节点信息传递给词级节点, 并且权重不共享. 同时, 我们使用了前面提到的异构图的 3 种边. 这两种方法使用不同的学习权重, 可以非自回归地更新每个节点的信息. 因为意图检测是粗粒度的句级分类任务, 槽位填充是相对细粒度的词级分类任务, 这两种任务信息是具有相互关联的不同粒度信息. 因此, 构造异构图可以显式区别这两种类型的节点信息, 并且在特定特征空间中学习两种类型节点的非重叠特征知识, 从不同粒度的状态信息中学习其潜在特征并进行集成.

表 2 单一组件消融实验 (%)

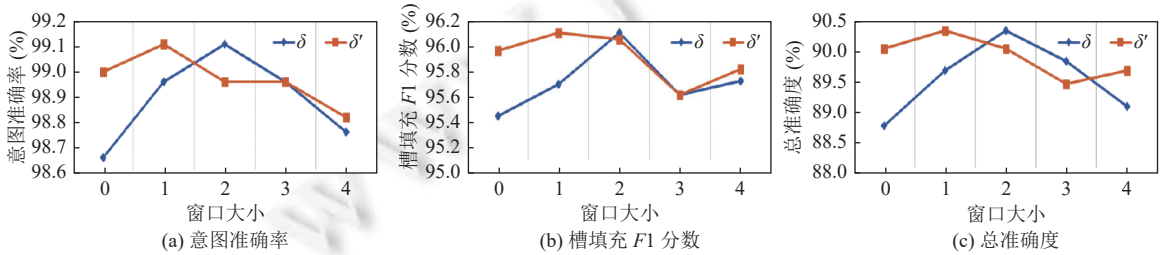
消融实验	ATIS			SNIPS		
	Intent Acc	Slot $F1$	Overall Acc	Intent Acc	Slot $F1$	Overall Acc
HcoSG	**97.98**	**96.10**	**88.02**	**99.11**	**96.11**	**90.35**
两级意图解码器消融实验	97.53	95.57	87.34	98.85	95.62	89.43
仅使用自注意力机制	97.76	96.00	87.68	98.71	95.90	89.73
仅使用图注意力网络	97.42	95.89	87.23	98.85	95.88	89.88
窗口机制消融实验	97.74	95.82	87.12	98.57	95.82	89.43
意图感知槽填充解码器消融实验	97.54	95.67	86.90	98.71	95.65	89.29

此外, 这部分结构最终用于槽位填充任务. 如前所述, 给定两个特定任务, 我们的模型可以学习路径中节点及其邻居之间的注意力值. 一些对不同特定任务有用的重要邻居往往具有更大的注意力值. 词级节点的更新过程包含两种抽象边, 即意图-槽连接边和槽-槽连接边. 由于两种类型的边使用不同的学习权重, 因此可以将词级槽信息与更丰富的句级语义信息相结合.

(3) 窗口机制消融实验

本部分研究是进行窗口机制消融实验, 可以认为去除了自注意力机制中 $f_{\text{window}}(\cdot)$ 函数的功能. 表 2 表明窗口机制是为实现更好性能所必需的. 从结果可以看出, 所有的评价指标都下降了, 这意味着我们忽略了槽位信息的时间顺序和位置, 即忽略了“O”标签局部出现的特性以及“B-”“I-”成对出现的局部性. 由于槽位填充和意图检测是两个相关性很强的任务, 利用槽的局部连续性可以更好地解码节点特征, 也可以更好地捕获“B-”“I-”槽信息来指导意图识别任务.

我们还在 SNIPS 数据集上进行了窗口大小实验, 以验证不同窗口大小对这两个任务的影响. 我们控制单个变量并通过仅调整其中一个窗口的大小来执行实验. 结果如图 3 所示. 我们的模型通过将异构单元中的自注意力窗口大小 δ 设置为 2 并将意图感知槽填充解码器中的自注意力窗口大小 δ' 设置为 1 来实现最佳性能. 我们可以看到, 槽位填充 $F1$ 分数、意图检测准确率和整体准确率都随着窗口大小的增加或减少而下降, 并且当异构结构中的自注意力的窗口大小 δ 设置为 0 时下降尤为明显. 这是因为当窗口大小设置为 0 时, 这意味着当前节点不与上下文相邻节点进行交互, 仅单纯地根据模型进行自适应更新, 丢失了大量的上下文信息和时序信息. 同时, 我们分析一个话语的槽标签是具有一定的局部性的, 一个话语的单词也是与时间顺序有关联的. 窗口机制可以准确地表示话语的标记, 窗口大小的灵活调整可以鲁棒性地适应槽标签和特定数据集的局部连续性.



(a) 意图准确率 (b) 槽填充 $F1$ 分数 (c) 总准确度

图 3 SNIPS 数据集上不同窗口大小的意图准确度、槽位填充 $F1$ 分数和整体准确度实验

(4) 意图感知槽填充解码器消融实验

为了验证意图感知槽填充解码器的有效性, 我们移除了这个解码器, 而是直接将异构单元的词级槽嵌入表示用于最终的槽位填充预测任务. 在表 2 中, 我们可以观察到 ATIS 和 SNIPS 数据集的整体准确率显著下降了 1.12% 和 1.06%, 这证明了意图感知槽填充解码器的重要性和有效性. 我们认为这是因为每个词级槽节点集成了所有意图标签, 并没有真正发挥预测意图的引导作用, 导致过度关注一些对最终槽位填充预测有副作用的意图节点. 这表明该组件充分且没有过度利用最终意图检测的信息来执行槽位填充任务, 同时最终的自注意力机制可以更好捕捉应用于最终任务的槽位信息, 可以提高口语理解性能.

4.6 多组件消融实验与分析

在第 4.5 节中, 我们依次剔除了 HcoSG 中的各个关键组件, 已验证了单一组件的有效性. 在本节中, 我们将进一步进行多组件的消融实验, 实验过程我们将移除多个核心组件, 其旨在探究各个组件之间的相互作用和其性能的提升是否重叠.

(1) FSD+Self Attention/FSD+GAT

针对本部分的多组件消融实验, 我们仅保留了模型中两级意图解码器中的第 1 级解码器 (first-stage decoder, FSD), 意图节点信息采用随机初始化的形式, 同时我们仍将异构交互注意力层中的异构结构替换为同构结构, 并且取消意图感知槽填充解码器. 如表 3 所示, FSD+Self Attention 指模型中仅包含第 1 级意图解码器和自注意力同构结构, 该同构结构的输入形式依旧是所有意图标签嵌入和话语单词嵌入的拼接形式. 相同地, FSD+GAT 中的同构结构使用图注意力网络. 结果所示, 3 个指标均有显著的下降. 我们认为在这样的设置中, 随机初始化意图嵌入信息不是最优的, 针对该模型而言, 意图信息的随机初始化的不确定性和不准确性会直接影响槽位单词嵌入表示的学习, 导致其学习到过多无用的信息. 同时在槽位填充任务的学习过程中, 意图节点信息会不断优化更新, 但是其并未直接影响到意图的预测, 这导致了这两个任务并未完全准确地进行交互学习. 针对 ID 和 SF 任务, 同第 4.5 节结论一致, 异构结构信息可以自适应地捕获不同类型和不同粒度的节点信息, 因为异构结构模式限定了对象集合以及对象间关系的类型约束, 这些约束使得异构信息网络具有半结构化的特点, 引导着意图语义和话语单词语义之间的知识共享. 本文在异构结构中设置的 3 种类型的边, 其链接的异质性考虑了不同对象之间的类型关系, 并建模它们的交互过程. 意图感知槽填充解码器具有遴选预测意图和加强特定意图下槽位填充任务的作用, 使得槽位词嵌入信息与最准确的意图进行交互实现最终的槽位值预测. 综上, 我们的模型中的核心组件两级意图解码器、异构交互注意力结构和意图感知槽填充解码器发挥着其各自的作用并提升模型的整体性能, 分别执行基于先验知识的意图标签嵌入生成、不同粒度信息之间的建模交互和预测意图指导槽位信息填充的作用.

表 3 多组件消融实验 (%)

消融实验	ATIS			SNIPS		
	Intent Acc	Slot *F*1	Overall Acc	Intent Acc	Slot *F*1	Overall Acc
HcoSG	**97.98**	**96.10**	**88.02**	**99.11**	**96.11**	**90.35**
FSD+Self Attention	96.78	95.57	86.23	98.08	94.22	87.56
FSD+GAT	96.35	95.52	86.04	97.85	94.45	87.80
FSD+Intent-Aware SFD	95.93	95.42	85.81	97.69	93.87	86.63

(2) FSD+Intent-aware SFD

本部分我们直接采用第 1 节意图解码器 (FSD) 和意图感知槽填充解码器 (intent-aware slot filling decoder, Intent-aware SFD) 来进行 ID 和 SF 任务的预测工作, 即第 1 级意图解码器进行话语整体意图的预测, 其选择预测出的经过初始化的意图嵌入表示进行最终槽位填充的指导. 结果如表 3 所示, 模型此时性能达到最低. 我们分析频繁预测出的意图标签的嵌入表示可以多次在意图感知槽填充解码器中更新, 以达到相对准确的意图嵌入表示, 但是很少被预测出的意图标签其嵌入表示接近其初始化表示, 其中并未包含丰富的语义信息并且含有大量噪声, 这

将恶化最终的槽填充任务性能, 同时反向作用于以影响第 1 级意图解码器的意图预测任务性能. 因此利用先验知识生成的意图标签嵌入表示可以更鲁棒性地指导槽位填充任务.

4.7 可视化

为了更好地证明和理解我们提出的模型中的异构交互单元可以有效地利用异构结构中的不同粒度信息进行槽位嵌入信息的学习, 我们将一句对话和所有意图标签信息进行可视化操作, 如图 4 所示, 话语样例取自 SNIPS 数据集. 横坐标为话语单词序列, 纵坐标为 7 个意图标签. 图中表示异构结构中槽位词节点对所有意图标签节点的关注程度, 颜色越深其关注程度越强. 基于图中的对话“please play me a popular track from 1984”我们可以清楚地看到对话中的槽位嵌入信息的注意力权重成功地集中在正确的意图“Play music”上, 这意味着我们的异构单元可以准确捕捉到不同粒度信息之间的共享知识, 并正确利用意图标签指导槽位填充任务. 同时, 综合上述实验结果显示, 意图感知槽填充解码器对异构交互注意力结构具有指导意义, 因为其使得意图标签嵌入表示可以充分利用槽位信息进行特征学习, 以达到准确表达意图的目的.

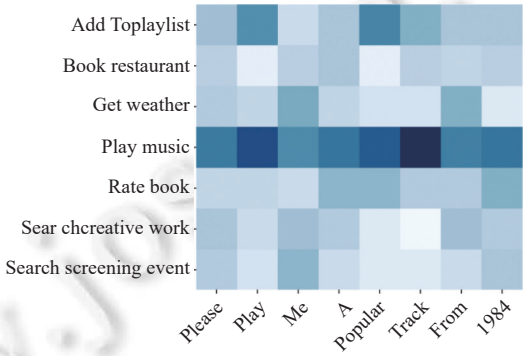


图 4 可视化

4.8 预训练模型的影响

最后, 我们还探讨了预训练模型 (BERT)[36]对我们框架的影响. 在本节中, 我们将嵌入编码器替换为 BERT 基础模型, 并与我们的模型保持相同的参数和组件. 我们对这两个数据集进行了实验, 实验结果如前文表 1 所示. 从实验结果可以看出, 我们的模型在应用预训练模型的同时展示了新的最先进的性能. 我们将其归因于预训练模型可以提供丰富的语义特征, 以提高两个下游任务的更好分类性能.

5 结 论

在本文中, 我们提出了一种用于联合意图检测和槽位填充的异构交互结构, 它充分捕捉了两个相关任务的异构信息中语义信息的复杂结构和丰富性. 此外, 我们在自注意力机制中采用了窗口机制, 可以充分利用词级槽位信息的局部连续性, 准确提取词级特征. 在两个公共数据集上的实验结果证明了我们所提出框架的有效性, 并且该框架实现了最先进的性能. 此外, 我们所提出的模型与预训练模型 BERT 相结合的效果可以使性能进一步提升.

References:

[1] Young S, Gašić M, Thomson B, Williams JD. Pomdp-based statistical spoken dialog systems: A review. Proc. of the IEEE, 2013, 101(5): 1160–1179. [doi: 10.1109/JPROC.2012.2225812]

[2] Tur G, De Mori R. Spoken Language Understanding: Systems for Extracting Semantic Information from Speech. New York: John Wiley & Sons, Ltd., 2011. [doi: 10.1002/9781119992691]

[3] Ni JJ, Young T, Pandelea V, Xue FZ, Cambria E. Recent advances in deep learning based dialogue systems: A systematic survey. Artificial Intelligence Review, 2023, 56(4): 3055–3155. [doi: 10.1007/s10462-022-10248-8]

[4] Haihong E, Niu PQ, Chen ZF, Song MN. A novel bi-directional interrelated model for joint intent detection and slot filling. In: Proc. of

the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 5467–5471. [doi: 10.18653/v1/P19-1544]

[5] Ramshaw LA, Marcus MP. Text chunking using transformation-based learning. In: Armstrong S, Church K, Isabelle P, Manzi S, Tzoukermann E, Yarowsky D, eds. Natural Language Processing Using Very Large Corpora. Dordrecht: Springer, 1999. 157–176. [doi: 10.1007/978-94-017-2390-9_10]

[6] Qin LB, Liu TL, Che WX, Kang BB, Zhao SD, Liu T. A co-interactive Transformer for joint slot filling and intent detection. In: Proc. of the 2021 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Toronto: IEEE, 2021. 8193–8197. [doi: 10.1109/ICASSP39728.2021.9414110]

[7] Zhang LH, Ma DH, Zhang XD, Yan XH, Wang HF. Graph LSTM with context-gated mechanism for spoken language understanding. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 9539–9546. [doi: 10.1609/aaai.v34i05.6499]

[8] Wu D, Ding L, Lu F, Xie J. SlotRefine: A fast non-autoregressive model for joint intent detection and slot filling. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. 1932–1937. [doi: 10.18653/v1/2020.emnlp-main.152]

[9] Liu YJ, Meng FD, Zhang JC, Zhou J, Chen YF, Xu JN. CM-Net: A novel collaborative memory network for spoken language understanding. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 1051–1060. [doi: 10.18653/v1/D19-1097]

[10] Qin LB, Xie TB, Che WX, Liu T. A survey on spoken language understanding: Recent advances and new frontiers. In: Proc. of the 30th Int'l Joint Conf. on Artificial Intelligence. Montreal: IJCAI.org, 2021. 4577–4584.

[11] Wang X, Ji Hy, Shi C, Wang B, Ye YF, Cui P, Yu PS. Heterogeneous graph attention network. In: Proc. of the 2019 World Wide Web Conf. San Francisco: Association for Computing Machinery, 2019. 2022–2032. [doi: 10.1145/3308558.3313562]

[12] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010

[13] Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: ICIR, 2018.

[14] Shi C, Li YT, Zhang JW, Sun YZ, Yu PS. A survey of heterogeneous information network analysis. IEEE Trans. on Knowledge and Data Engineering, 2017, 29(1): 17–37. [doi: 10.1109/TKDE.2016.2598561]

[15] Hemphill CT, Godfrey JJ, Doddington GR. The ATIS spoken language systems pilot corpus. In: Proc. of the 1990 Workshop on Speech and Natural Language. Hidden Valley: Association for Computational Linguistics, 1990. 96–101. [doi: 10.3115/116580.116613]

[16] Coucke A, Saade A, Ball A, Bluche T, Caulier A, Leroy D, Doumouro C, Gisselbrecht T, Caltagirone F, Lavril T, Primet M, Dureau J. Snips voice platform: An embedded spoken language understanding system for private-by-design voice interfaces. arXiv:1805.10190, 2018.

[17] Haffner P, Tur G, Wright JH. Optimizing svms for complex call classification. In: Proc. of the 2003 IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing. Hong Kong: IEEE, 2003. I-632–I-635. [doi: 10.1109/ICASSP.2003.1198860]

[18] Raymond C, Riccardi G. Generative and discriminative algorithms for spoken language understanding. In: Proc. of the 8th Interspeech Annual Conf. of the Int'l Speech Communication Association. Anvers: HAL, 2007.

[19] Deng L, Tur G, He XD, Hakkani-Tur D. Use of kernel deep convex networks and end-to-end learning for spoken language understanding. In: Proc. of the 2012 IEEE Spoken Language Technology Workshop (SLT). Miami: IEEE, 2012. 210–215. [doi: 10.1109/SLT.2012.6424224]

[20] Tur G, Deng L, Hakkani-Tür D, He XD. Towards deeper understanding: Deep convex networks for semantic utterance classification. In: Proc. of the 2012 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Kyoto: IEEE, 2012. 5045–5048. [doi: 10.1109/ICASSP.2012.6289054]

[21] Ravuri S, Stolcke A. Recurrent neural network and lstm models for lexical utterance classification. In: Proc. of the 16th Annual Conf. of the Int'l Speech Communication Association. Dresden: ISCA, 2015. 135–139. [doi: 10.21437/Interspeech.2015-42]

[22] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

[23] Wu CS, Hoi SCH, Socher R, Xiong CM. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. 917–929. [doi: 10.18653/v1/2020.emnlp-main.66]

[24] Yao KS, Zweig G, Hwang MY, Shi YY, Yu D. Recurrent neural networks for language understanding. In: Proc. of the 2013 Interspeech. Lyon: ISCA, 2013. 2524–2528.

[25] Mesnil G, He XD, Deng L, Bengio Y. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In: Proc. of the 2013 Interspeech. Lyon: ISCA, 2013: 3771–3775.

[26] Mesnil G, Dauphin Y, Yao KS, Bengio Y, Deng L, Hakkani-Tur D, He XD, Heck L, Tur G, Yu D, Zweig G. Using recurrent neural networks for slot filling in spoken language understanding. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2015, 23(3): 530–539. [doi: 10.1109/TASLP.2014.2383614]

[27] Coope S, Farghly T, Gerz D, *et al*. Span-ConveRT: Few-shot span extraction for dialog with pretrained conversational representations. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 107–121. [doi: 10.18653/v1/2020.acl-main.11]

[28] Zhang XD, Wang HF. A joint model of intent determination and slot filling for spoken language understanding. In: Proc. of the 25th Int'l Joint Conf. on Artificial Intelligence. New York: AAAI Press, 2016. 2993–2999.

[29] Liu B, Lane I. Attention-based recurrent neural network models for joint intent detection and slot filling. In: Proc. of the 2016 Interspeech. San Francisco: ISCA, 2016. 685–689. [doi: 10.21437/Interspeech.2016-1352]

[30] Goo CW, Gao G, Hsu YK, Huo CL, Chen TC, Hsu KW, Chen YN. Slot-gated modeling for joint slot filling and intent prediction. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 2 (Short Papers). New Orleans: Association for Computational Linguistics, 2018. 753–757. [doi: 10.18653/v1/N18-2118]

[31] Li CL, Li L, Qi J. A self-attentive model with gate mechanism for spoken language understanding. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 3824–3833. [doi: 10.18653/v1/D18-1417]

[32] Qin LB, Che WX, Li YM, Wen HY, Liu T. A stack-propagation framework with token-level intent detection for spoken language understanding. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 2078–2087. [doi: 10.18653/v1/D19-1214]

[33] Wang JX, Wei K, Radfar M, Zhang WW, Chung C. Encoding syntactic knowledge in transformer encoder for intent detection and slot filling. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2021. 13943–13951.

[34] Kim Y. Convolutional neural networks for sentence classification. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Doha: Association for Computational Linguistics, 2014. 1746–1751. [doi: 10.3115/v1/D14-1181]

[35] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2015.

[36] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2018. 4171–7186. [doi: 10.18653/v1/N19-1423]

张启辰(1993－), 男, 博士生, 主要研究领域为自然语言处理, 对话系统.

李静梅(1964－), 女, 博士, 教授, 博士生导师, 主要研究领域为自然语言处理, 大数据, 云计算.

王帅(1998－), 男, 硕士生, 主要研究领域为自然语言处理, 对话系统.