计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.181202269

基于上下文信息的口语意图检测方法

徐 扬 王建成 刘启元 李寿山
苏州大学计算机科学与技术学院 江苏 苏州 215006
(yxu2017@stu.suda.edu.cn)

摘 要 近年来,随着人工智能的发展与智能设备的普及,人机智能对话技术得到了广泛的关注。口语语义理解是口语对话系统中的一项重要任务,而口语意图检测是口语语义理解中的关键环节。由于多轮对话中存在语义缺失、框架表示以及意图转换等复杂的语言现象,因此面向多轮对话的意图检测任务十分具有挑战性。为了解决上述难题,文中提出了基于门控机制的信息共享网络,充分利用了多轮对话中的上下文信息来提升检测性能。具体而言,首先结合字音特征构建当前轮文本和上下文文本的初始表示,以减小语音识别错误对语义表示的影响;其次,使用基于层级化注意力机制的语义编码器得到当前轮和上下文文本的深层语义表示,包含由字到句再到多轮文本的多级语义信息;最后,通过在多任务学习框架中引入门控机制来构建基于门控机制的信息共享网络,使用上下文语义信息辅助当前轮文本的意图检测。实验结果表明,所提方法能够高效地利用上下文信息来提升口语意图检测效果,在全国知识图谱与语义计算大会(CCKS2018)技术评测任务2的数据集上达到了88.1%的准确率(Acc值)和88.0%的综合正确率(F1值),相比于已有的方法显著提升了性能。

关键词:口语语义理解;意图检测;上下文信息;门控神经网络
中图法分类号 TP391

Intention Detection in Spoken Language Based on Context Information

XU Yang,WANG Jian-cheng,LIU Qi-yuan and LI Shou-shan
School of Computer Science & Technology,Soochow University,Suzhou,Jiangsu 215006,China

Abstract In recent years,with the development of artificial intelligence and the popularization of smart devices,human-computer intelligent dialogue technology has received extensive attention. Spoken language understanding is an important task dialogue system,and spoken language intention detection is a key technology in spoken language understanding. Due to complex language phenomena such as semantic missing,frame representation and intent conversion in multiple rounds of dialogue,the intent detection task for spoken language is very challenging. In order to solve the above problems,a gated mechanism based information sharing neural network method was proposed in this paper,which can take advantages of contextual information in multiple rounds of dialogue to improve detection performance. Specifically,first the current round text and context text initial representation are constructed in combination with the phonetic features to reduce the impact of speech recognition errors on semantic representation. Secondly,a semantic encoder based on hierarchical attention mechanism is used to obtain deep semantic representations of the current round and contextual text,including multi-level semantic information from word to sentence to multiple rounds of text. Finally,the gated mechaniam based information sharing neural network is constructed to use the context semantic information to help the intent detection of the current round of text. The experimental results show that the proposed method can effectively use context information to improve the detection of spoken language intentions,and achieves 88.1% accuracy and 88.0% F1 value in dataset of CCKS2018 shared task-2,which is significantly improved performance compared with the existing methods.

Keywords Spoken language understanding,Intent detection,Context information,Gated neural network

1 引言

近年来,随着各类智能设备的普及,人机交互的途径逐渐变得多样化与智能化。各种交互途径中,对话是人类最自然的交流方式,因此口语对话系统有着广阔的研究前景与应用价值[1-2]。简要地,口语对话系统是一个能够理解用户口语表述的语义并做出相应反馈的智能交互系统。其中,口语语义理解是对话系统中的一项重要任务,只有准确地理解用户语义才能使系统做出精准的反馈。因此,口语语义理解的相关研究越来越受到人们的关注。

到稿日期:2018-12-05 返修日期:2019-04-24 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:国家自然科学基金(61331011,61375073)
This work was supported by the National Natural Science Foundation of China (6133101,61375073).
通信作者:李寿山(lishoushan@suda.edu.cn)

口语语义理解任务一般被看作两个子任务(意图检测与语义槽填充)的结合。具体地,对一句用户话语的理解过程分为两步,首先通过意图检测得到用户的意图,再通过槽值抽取得到与该意图相关的语义槽值。例如,对用户话语"请帮我点播一首周杰伦的七里香"的语义理解过程可以表述为:首先检测意图为点播歌曲,接着抽取与点播歌曲相关的语义槽(演唱者:"周杰伦",歌曲名:"七里香"),在得到语义理解后,对话系统就能根据语义理解结果向用户做出正确的反馈,推送播放链接或者直接播放相应的歌曲。由此可见,意图检测是实现口语语义理解的关键技术,本文主要关注面向对话系统的口语意图检测任务。

口语意图检测任务通常被看作特定交互场景下的文本分类任务,将对话过程中的口语文本分类到所属的意图类别。

以往的研究表明,常规的文本分类模型已经能够在口语意图检测任务上取得不错的效果,但是想要获得更好的检测效果,不能简单地套用通用领域的文本分类模型,与常规的文本分类任务相比,口语意图检测任务面临着更多挑战。首先,口语意图检测通常使用自然语音的语音解析结果作为输入,解析过程中不可避免地会出现错字、噪音等错误,如表 1 中的示例 1,将"唱"错误解析为近音字"张",解析错误会对意图检测的效果产生负面影响;其次,与书面文本相比,口语文本常常充满着语义缺失、颠倒、重复等现象[4],如表 1 中的示例 2、示例 3,其中"父亲"可以是指点播歌曲《父亲》,也可以是拨通联系人父亲,需要结合上下文语境才能补全缺失的语义信息;最后,在实际交互场景下用户与系统通常会进行多轮交互,句子级别的语义框架无法完整地保留口语文本的语义信息[5],同时如表 1 中的示例 4 所示,多轮交互过程中用户的意图可能会发生多次转变,需要针对性地提出面向多轮交互的语义表示框架,以解决语义表示与意图转变的问题。

表 1 多轮话语示例

Table 1 Example of multiple rounds of utterance

当前轮话语	历史轮话语	意图
"张(唱)首歌"	"你这机器人会干点啥"	点歌
"父亲"	"选取联系人"	通话
"父亲"	"点一首筷子兄弟的歌"	点歌
"算了放歌吧"	"打开桌面文件"	点歌

口语意图检测任务中存在上述问题,因此本文提出了一种基于门控机制的信息共享神经网络方法,该方法能够充分利用多轮对话中的上下文信息来提升检测性能,同时能降低语音解析错误对意图检测结果的影响。具体而言,在口语文本的嵌入表示模块中,结合字音特征的文本嵌入方法能减小语音解析错误对意图检测的负面影响;在语义表示模块中,基于层级化注意力机制的多级语义编码器能得到当前轮和上下文文本的深层语义表示,包含由字到句再到多轮文本的多级语义信息;在意图检测模块中,在多任务学习框架中引入了门控机制构建基于门控机制的信息共享网络,使用上下文语义信息辅助当前轮文本的意图检测。为了验证该方法的有效性,本文使用 2018 年全国知识图谱与语义计算大会(CCKS2018)中评测任务 2 的数据进行实验。该评测任务面向音乐领域的用户语义理解,给出了用户对智能终端的多轮命令文本,要求解析最后一轮命令文本的语义。在各参赛队提交的方法中,本文提出的方法在意图检测模块取得了第 1 名的成绩。

本文第 2 节介绍了口语意图检测的一些相关研究;第 3 节详细描述了本文针对口语意图检测任务提出的基于门控机制的信息共享网络;第 4 节介绍了实验设置并对实验结果进行了分析;最后总结全文并展望下一步工作。

2 相关工作

面向对话系统的口语意图检测经过多年的研究已经颇多成果[3]。早期,研究者主要通过预先制订的规则模板对话语进行句法分析或语义分析来理解口语语义[6-8],但是由于口语文本的不规范性,给规则模板的制订带来了很多困难,同时制定规则的过程十分繁复且具有主观性,因此此类方法没有得到大规模的应用。随着 20 世纪末以航空旅行信息系统语料库(ATIS)为代表的各种面向口语语义理解的语料库的发布,基于统计学习的方法逐渐占据了研究主流[9],Minker 等[10]采用的隐马尔可夫模型、Haffne 等[11]使用的支持向量机模型以及 Schapires 等[12]使用的集成学习模型都在口语意图检测任务上取得了良好的效果。

近年来,基于深度神经网络的深度学习模型在各项自然语言处理任务上都取得了优越的性能。在口语意图检测方面,最早期的尝试为 Sarikaya 等[13]提出的深度置信网络(Deep Belief Network,DBN)以及 Tur 等[14]提出的深度语义网络(Deep Convex Networks,DCNs),这两种模型在当时都取得了优于浅层机器学习模型的性能。随后,Hashemi 等[15]首先使用了卷积神经网络(Convolutional Neural Network,CNN)对用户话语进行语义编码,将得到的语义编码作为意图分类的特征。同样地,循环神经网络(Recurrent Neural Network,RNN)常被用于序列建模,因此在后续的研究中各种以 RNN 为基础的模型在口语意图检测任务中得到了广泛的应用。Ravuri 等[16]提出了基于长短期记忆网络(Long Short-Term Memory,LSTM)的意图检测模型;Liu 等[17]使用了基于注意力机制的双向长短期记忆网络(Bi-LSTM)进行口语意图检测和槽值填充;Mauajama 等[18]使用了 CNN 与 RNN 的深度集成模型检测口语意图。

近期,口语意图检测的相关研究主要关注如何更好地对多轮交互场景进行建模,即在对当前轮文本进行意图检测时更好地利用上下文信息。Young 等[19]提出了 CNN-LSTM 模型,通过 LSTM 和 CNN 分别对上下文语境和当前轮话语进行语义表示,并将两种表示拼接得到更高层的语义表示,以提升意图检测效果。Xie 等[20]提出了层级化的语义编码模型,使用 Bi-LSTM 在句子级别和文档级别两个层级对多轮对话进行多级语义编码,得到了包含上下文语境信息的全局语义表示,并用于意图检测。

不同于已有的研究,本文提出的方法具有以下创新点与贡献:1)针对口语文本语音解析错误问题,提出了一种结合字音特征的文本嵌入方式,能够减小解析错误对意图检测的负面影响;2)针对口语文本语义缺失问题,构建了基于层级化注意力机制的多级语义编码器,能够获得当前轮和上下文文本的深层语义表示,并利用上下文信息对当前轮文本进行语义补充;3)针对多轮对话中的语义转变问题,在多任务学习框架

中引入门控机制来构建信息共享网络，根据上下文与当前轮文本的语义相关度灵活地进行信息共享以提升意图检测效果。实验结果表明，相比于已有的研究，该方法的性能得到了显著提升。

3　信息共享学习方法

图 1 给出了本文提出的基于门控机制的信息共享网络，该神经网络模型分为以下 3 个部分。

(1)口语文本嵌入表示模块：结合字音特征将文本映射为固定维度的向量，并将其作为用户话语文本的初始表示。

(2)多级语义表示模块：通过基于层级化注意力机制的多级语义编码器，将口语文本的初始表示编码为当前轮和上下文文本的深层语义表示。

(3)意图检测模块：构建信息共享网络，通过信息共享融合当前轮与上下文文本的语义表示，以提升意图检测的效果。

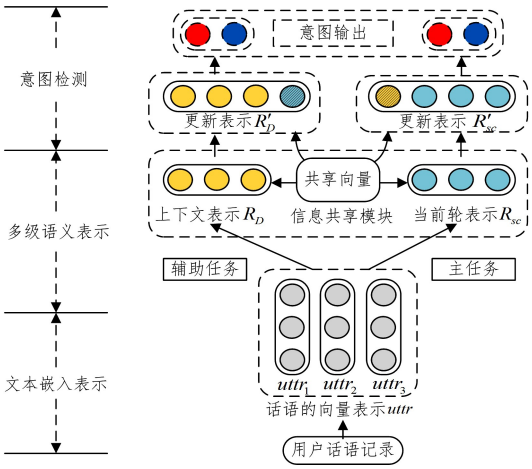

图 1　基于门控机制的信息共享神经网络

Fig. 1　Gated mechanism based information sharing neural network

3.1　口语文本嵌入表示

由于口语文本包含许多语音解析错误与冷僻词语，分词处理效率不高，因此本文不对口语文本做分词处理，直接将其视为基于字的文本序列 S 进行处理：

$$S=[w_1,w_2,\cdots,w_t,\cdots,w_T] \tag{1}$$

其中，$w_t\in\mathbb{R}^{|V|}$ 表示序列中第 t 个字在字符表 V 中的 one-hot 表示，T 表示文本序列长度，$|V|$ 表示字符表的大小。

在自然语言处理领域中，通常使用稠密固定维度的向量对字或者词进行嵌入表示，并将其作为神经网络的输入[21]，尤其是自 Mikolov 等[22]提出 word2vec 模型以来，文本的嵌入表示成为了各项自然语言处理任务的基础模块。因此，本节使用 word2vec 模型训练基于字的口语文本序列来得到字向量矩阵，在得到字向量矩阵后，字符 w_t 可以映射为固定维度的字向量 x_t：

$$x_t=W_e\cdot w_t \tag{2}$$

其中，$W_e\in\mathbb{R}^{d\times|V|}$ 表示字向量矩阵，d 表示字向量的维度。

需要注意的是，口语文本中往往不可避免地存在语音解析错误，同一个字会被解析为不同的同音字或近音字，因此只使用单一的字向量矩阵进行嵌入表示，会把实际相同的语义映射为不同的表达，给意图检测带来负面影响。

近年来很多研究表明，在处理中文语料时使用拼音特征能够很好地表示语义。拼音由包含中文字符字音信息的拉丁字母组成，而使用汉字拼音作为特征最直接的优势是能够拉丁化表示中文字符。经过拉丁化表示后的中文字符就可以使用面向英文语料(如提取词语前缀后缀表示等)的处理手段来增加特征以表示语义[23]。其次，中文语料中的语音识别错误通常表现为将同一个字解析为韵母相同的近音字或同音字，如以下两种解析结果：

正确："唱(chang)首周杰(jie)伦的歌"

实际："张(zhang)首周捷(jie)伦的歌"

正解与误解为语义不相关的汉字，但是实际对应的韵母相同，因此可以选取拼音表示的后缀——韵母来作为附加的字音特征，通过加入字音特征能够拉近被错误解析的字符在文本表示向量空间上与正确解析的几何距离，从而减小语音解析错误对文本表示的负面影响[24]。

因此，本节在使用字向量矩阵 $W_e\in\mathbb{R}^{d\times|V|}$ 的同时，还使用字音向量矩阵 $W_e'\in\mathbb{R}^{d'\times|V'|}$ 来映射语义：

$$x_t'=W_e'\cdot w_t \tag{3}$$

在得到对应的字音向量之后，结合字音特征的语义嵌入就可以表示为：

$$X_t=[x_t,x_t'] \tag{4}$$

其中，$[,]$表示向量的拼接操作，V' 表示字音表，d'表示字音向量的维度，x'表示字音向量，X_t 表示最终的字向量表示。口语文本序列 S 经过嵌入表示为向量序列 $uttr$：

$$uttr=[X_1,X_2,\cdots,X_t,\cdots,X_T] \tag{5}$$

3.2　多轮对话的多级语义表示

在多轮对话的场景下，需要对当前轮文本与上下文文本进行语义编码表示。可以将多轮话语直接首尾拼接成一个序列，使用句子级别的语义模型来对其进行处理，但是研究表明[25]，随着对话轮次的增加，拼接序列增长，单一的句子级别模型无法高效地学习到超长序列中的远程依赖。

本节提出了基于层级化注意力机制的多级语义编码器，其在句子级别和文档级别两个层面编码多轮对话，能够同时得到当前轮与上下文文本的语义表示，具体结构如图 2 所示。

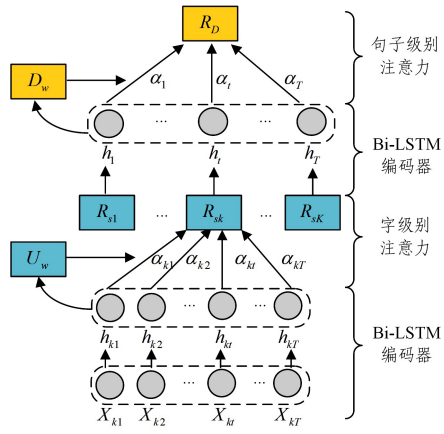

图 2　基于层级化注意力机制的多级语义编码器

Fig. 2　Hierarchical attention based multi-level semantic encoder

3.2.1　句子级别的单轮语义编码

多轮对话由多个单轮话语组成，每个单轮话语可以看作句子级别的文本序列来处理。如图 2 所示，在本文提出的多级语义编码器中，在句子级别对所有单轮对话使用 Bi-

LSTM[26]进行编码，具体过程如下：

$$\overrightarrow{h_t}=\overrightarrow{LSTM}(x_t) \tag{6}$$

$$\overleftarrow{h_t}=\overleftarrow{LSTM}(x_t) \tag{7}$$

$$h_t=[\overrightarrow{h_t},\overleftarrow{h_t}] \tag{8}$$

其中，$\overrightarrow{h_t}$ 与 $\overleftarrow{h_t}$ 分别为文本序列中 t 时刻的输入通过前向和反向 LSTM 得到的状态表示，h_t 表示文本序列中 t 时刻的输入通过 Bi-LSTM 得到的状态表示。

注意力机制[27]是在神经机器翻译任务下被提出的一种结构，一经提出便迅速地应用于其他各项自然语言处理任务并获得了很好的效果。将其应用于文本语义表示任务时，能够学习得到文本序列中各部分对文本深层语义表示的贡献权重，因此在处理单轮话语时使用注意力机制学习序列中各个输入的贡献权重，以得到更准确的深层语义表示，过程如下：

$$u_{kt}=\tanh(W_w \cdot h_{kt}+b_w) \tag{9}$$

$$\alpha_{kt}=\frac{\exp(u_{kt}^{\mathrm{T}} \cdot u_w)}{\sum_t \exp(u_{kt}^{\mathrm{T}} \cdot u_w)} \tag{10}$$

$$R_{sk}=\sum_t \alpha_{kt} \cdot h_{kt} \tag{11}$$

其中，h_{kt} 表示第 k 轮话语的 t 时刻的状态，h_{kt} 通过参数矩阵 W_w 和 b_w 得到中间状态表示 u_{kt}。再用此中间状态 u_{kt} 与背景语境向量 u_w 的相似度来表示其注意力权重 α_{kt}。最后得到第 k 轮文本序列的编码向量 R_{sk}，其表示中间状态向量 h_{kt} 与其注意力权重 α_{kt} 的加权和。

3.2.2 文档级别的多轮语义编码

上下文文本由多轮对话文本组成，不同于将多轮话语拼接成一个长序列的做法，本文将多轮对话文本看作由多个单轮话语组成的文档级文本进行处理，在得到句子级别的单轮语义表示 R_s 后，将多轮话语 D 看作由多个单轮语义表示组成的序列：

$$D=[R_{S1},R_{S2},\cdots,R_{Sk},\cdots,R_{SK}] \tag{12}$$

其中，R_{Sk} 表示第 k 轮命令的语义表示，如图 2 所示，在由多个单轮语义表示组成的多轮话语序列上，使用与单轮编码一致的基于注意力机制的 Bi-LSTM 网络进行编码：

$$h_k=Bi\text{-}LSTM(R_{Sk}) \tag{13}$$

$$R_D=\sum_k \alpha_k \cdot h_k \tag{14}$$

其中，h_k 是经过 Bi-LSTM 编码后，序列中的第 k 时刻的状态；R_D 是注意力机制得到的各轮话语的语义加权表示，将其作为多轮对话的深层语义表示。

3.3 基于信息共享的意图检测

由上述分析可知，对多轮对话中当前轮文本进行意图检测时需要结合上下文信息，相同的当前话语结合不同的上下文信息就指向不同的意图。不同于常规的向量拼接融合，本文使用多任务学习框架并结合上下文信息与当前轮的语义信息。多任务学习由 Caruna 等[28]首次提出，通过多个任务之间共享信息来提升模型的性能表现。现今，多任务学习框架在自然语言处理领域中得到了广泛的应用[28-29]。在意图检测模块中，通过多任务学习框架中的信息共享传递机制，可以有效地联合共享学习当前轮语义表示与上下文信息，使用上下文语义信息辅助当前轮文本的意图检测，解决语义缺失的问题。

基于此，本文提出了基于信息共享的意图检测方法。如图 3 所示，通过多任务学习框架同时学习两个任务：1）当前轮文本的意图检测任务，为框架中的主任务；2）上下文文本的意图检测任务，为辅助任务。主辅之间通过信息共享模块进行信息融合共享。

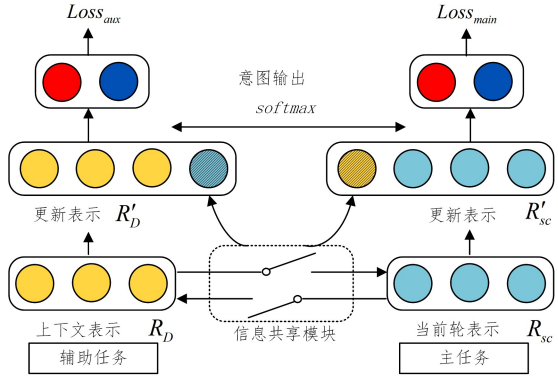

图 3 基于信息共享的意图检测模块

Fig. 3 Information sharing based intent detection model

将上下文信息与当前轮文本信息简单地融合无法应对意图转换的现象。当对话语义发生转变，历史语境与当前轮文本意图不同时，上下文信息的加入反而会成为当前话语意图检测的噪音，给意图检测带来负面影响，因此需要控制上下文信息与当前轮文本之间的信息共享程度。因此，本文考查了两种信息共享的机制：硬共享模式和基于门控机制的软共享模式。

3.3.1 硬共享模式

硬共享模式是各种多任务学习模型中常用的一种信息共享机制[30]，其核心思想是直接将多任务学习框架中各个任务的中间语义表示进行拼接，得到的高层语义表示包含了多个任务中的语义信息。将其应用到本文提出的框架时就表现为：首先主辅任务分别通过多级语义编码器得到当前轮文本和上下文文本的深层语义表示 R_{sc} 和 R_D，其次在信息共享模块中直接将在辅助任务中得到的上下文语义表示 R_D 作为共享信息加入到主任务中，以更新当前轮话语的语义表示 R_{sc}。具体表达式如下：

$$R'_{sc}=[R_{sc},R_D] \tag{15}$$

信息共享后，主任务中对当前轮话语的语义表示就更新为包含上下文信息的高层表示 R'_{sc}。

3.3.2 基于门控机制的软共享模式

通过硬共享模式可以将上下文信息融入到当前轮文本的语义表示中，然而在这种基于简单向量拼接的模式中，辅助任务无法学习到主任务的信息，也无法控制任务之间的信息共享程度。在上下文意图与当前轮话语意图不同时，使用硬共享模式得到的共享信息会成为对主任务的噪声。因此，本文提出了基于门控机制的软共享模式，在信息共享模块中使用双向门控机制替代单向的信息拼接以控制主辅任务之间的信息共享程度，当前轮语义与上下文语义之间的相关性系数 f_M 的计算式为：

$$f_M=\sigma(W_f \cdot [R_{sc},R_D]+b_f) \tag{16}$$

其中，σ 表示 sigmoid 函数，W_f 与 b_f 表示参数矩阵。通过式(16)，门控机制可以通过在主辅任务中得到的语义表示 R_{sc} 和 R_D，来学习当前轮语义与上下文语义之间的相关性系数

f_M，并将其作为信息共享模块中的共享权重，得到共享权重后共享信息表示为两个任务中的语义表示与共享权重的加权和。

$$R_{aux}=f_W\odot R_D \tag{17}$$

$$R_{main}=f_W\odot R_{sc} \tag{18}$$

其中，R_{aux} 表示辅助任务向主任务分享的信息，R_{main} 表示主任务向辅助任务分享的信息。

在门控机制下，共享信息的规模会随着当前轮对话与上下文的语义相关度的高低变化而增大或减小。当上下文与当前轮语义的相关度较低时，共享权重就会减小以减弱噪声信息对结果的负面影响；当上下文与当前轮语义的相关度较高时，共享权重就会增大以增大上下文信息与当前轮文本的信息共享规模，从而实现在充分共享信息的同时还能控制噪声信息对结果的影响。最终，主辅任务中各自的语义表示通过与共享信息进行拼接更新为了更全面的语义表示 R'_{sc} 与 $R_D{}'$。

$$R'_{sc}=[R_{sc},R_{aux}] \tag{19}$$

$$R_D{}'=[R_D,R_{main}] \tag{20}$$

更新后的高层语义表示通过 softmax 层得到当前轮文本与上下文文本属于各个意图的概率分布：

$$p_{main}=softmax(W_{sc}\cdot R'_{sc}+b_{sc}) \tag{21}$$

$$p_{aux}=softmax(W_D\cdot R_D{}'+b_D) \tag{22}$$

其中，p_{main} 与 p_{aux} 是当前轮文本与上下文文本在各个意图类别上的概率分布，W_{sc} 和 W_D 为 softmax 层的权重，b_{sc} 和 b_D 为 softmax 层的偏置。

3.4　训练与优化策略

在网络的训练过程中，本文采用 Adam 优化算法[30]更新网络中的参数，且主任务与辅助任务均使用交叉熵误差作为损失函数：

$$Loss(\hat{y},y)=-\sum_{k=1}^{N}\sum_{i=1}^{C}y_k^i\cdot\log\hat{y}_k^i \tag{23}$$

其中，y 是真实标签，$\hat{y}$ 是模型预测的概率，N 是训练样本总数，C 是意图类别的数目。同时将辅助任务与主任务的损失函数的加权和作为整体网络的损失函数：

$$Loss=Loss_{main}+w_{loss}Loss_{aux} \tag{24}$$

其中，w_{loss} 表示辅助任务损失函数 $Loss_{aux}$ 相对于主任务的损失函数 $Loss_{main}$ 的权重。

4　实验及结果分析

本节将系统地分析本文提出的方法在口语意图检测任务上的实际效果。

4.1　实验设置

本文实验基于 2018 年全国知识图谱与语义计算大会(CCKS2018)的评测任务 2 提供的数据，该评测面向多轮用户话语，要求判断其中最新一轮的用户话语是指向点歌的意图，可以将其看作一个二分类的意图检测任务。评测使用的数据来自人机对话系统中音乐领域以及非音乐领域的真实用户话语请求记录。CCKS2018 评测数据集[1]共包含 12000 组数据，每组数据由用户的 3 轮话语组成。实验将数据集以7∶1∶2的比例随机划分为训练集、验证集和测试集。

实验采用 word2vec[2]训练字向量与字音向量，使用 pypinyin[3]开源工具包提取文本字音。神经网络模型的搭建通过 Keras[4]框架实现，主要参数设置如表 2 所列，网络训练过程使用 NVIDIA TITAN Xp 显卡进行加速。对于实验结果的评估，本文使用准确率(Accuracy)和 F1 值作为衡量意图检测性能的评价指标。

表 2　神经网络中的主要参数设置

Table 2　Parameters setting in neural network

参数	数值
字向量维度	300
字音向量维度	300
LSTM 层输出维度	128
批大小	64
Dropout rate	0.5
辅助任务损失权重	0.25
迭代次数	20

4.2　实验结果

为了验证本文意图检测方法的有效性，将几种近年来在口语意图检测任务上具有代表性的工作进行比较，实验结果如表 3 所列。

表 3　本文方法与其他方法的比较

Table 3　Comparison of proposed method with other methods

方法	Accuracy	F1
DBN	0.676	0.615
SLU2	0.764	0.757
DEN	0.860	0.860
Attention-RNN	0.853	0.851
Hierarchical-LSTM	0.868	0.868
CNN+LSTM	0.871	0.869
ISNN	0.872	0.873
ISNN(+py)	0.874	0.874
Gated-ISNN	0.877	0.877
Gated-ISNN(+py)	**0.881**	**0.880**

DBN：它是由 Sarikaya 等[13]提出的最早应用在意图分类任务上的深度神经网络模型。

SLU2：它是 Williams 等[32]为第二届对话状态追踪挑战(DSTC2)开发的口语理解系统，使用基于决策树的模型并提取用户话语的 bi-gram 特征作为模型输入。

DEN：它是 Mauajama 等[18]提出的深度集成方法，同时使用 GRU，LSTM 以及 CNN 对用户话语进行编码，将得到的编码表示拼接为高层的语义表示。

Attention-RNN：它是 Wang 等[33]提出的基于注意力机制与 RNN 的意图检测模型。

Hierarchical-LSTM：它是由 Xie 等[20]在第六届对话状态追踪挑战(DSTC6)中提出的模型，模型中构建层级化的 Bi-LSTM 结构对多轮话语进行语义建模。

CNN+LSTM：它是由 Young 等[19]提出的口语语义理解模型，分别使用 CNN 与 LSTM 对用户话语和一定窗口长度的上文进行语义表示，将两种表示结合得到意图结果。

ISNN：它是本文提出的信息共享网络(Information Sharing Neural Network)，其中文本嵌入表示模块不结合字音特征且多任务学习框架内使用硬共享的信息共享模式。

1) http://www.ccks2018.cn

2) https://github.com/dav/word2vec

3) http:// http://pypinyin.mozillazg.com

4) http://Keras.io

ISNN(+py):它是本文提出的信息共享网络,其中文本嵌入表示模块结合字音特征且在多任务学习框架内使用硬共享的信息共享模式。

Gated-ISNN:它是本文提出的基于门控机制的信息共享网络(Gated Information Neural Network),其中文本嵌入表示模块不结合字音特征且在多任务学习框架内使用门控软共享的信息共享模式。

Gated-ISNN(+py):它是本文提出的基于多任务学习的口语意图检测模型,其中文本嵌入表示模块结合字音特征且在多任务学习框架内使用门控软共享的信息共享模式。

由表3可知,在准确率以及综合正确率这两项评价指标上,本文提出的4种基于信息共享的神经网络方法在所有方法中达到了最优的表现,说明在面向多轮对话的意图检测任务中,基于上下文信息共享的神经网络方法相对于已有方法具有更优越的性能表现。同时,通过进一步分析可以得到:首先,在文本表示模块中,使用结合字音特征的方法均比不使用的方法在各项指标上提高了0.5%~1%,说明这种结合字音特征的嵌入表示方法是有效的,能够减弱语音解析错误对意图检测的负面影响;其次,在意图检测模块中,使用门控软共享机制的方法比使用硬共享机制的方法在各项指标上提高了1%左右,说明在信息共享网络中引入门控机制来控制信息共享的规模能够显著提升意图检测的效果;最后,使用字音特征与双向门控机制的信息共享网络方法Gated-ISNN(+py)获得了88.1%的准确率以及88.0%的综合正确率,在所有的方法中获得了最佳的指标。

4.3 样例分析

为了更好地理解Gated-ISNN的优越性,本节通过表4中的两个样例做进一步的分析说明。

表4 本文方法的表现

Table 4 Performance of proposed method

样例一	
uttr1:"想听筷子兄弟的歌" uttr2:"筷子兄弟筷子兄弟,我要听筷子兄弟的歌" uttr3:"父亲"	
Attention-RNN ×(不点播歌曲)	CNN+LSTM √(点播歌曲)
ISNN √(点播歌曲)	Gated-ISNN √(点播歌曲)
样例二	
uttr1:"你会说话吗,你会干点啥,你会预测天气吗" uttr2:"帮我查查明天天气怎么样啊" uttr3:"算了算了还是给我放歌吧,一万次悲伤"	
Attention-RNN √(点播歌曲)	CNN+LSTM ×(不点播歌曲)
ISNN ×(不点播歌曲)	Gated-ISNN √(点播歌曲)

样例1中,多轮对话中最后一轮的"父亲"有语义缺失的现象,并未明确地指向音乐意图,若不结合前两轮的信息无法判别其真实意图。因此,基于句子级别模型的方法(Attention-RNN)将此样例判断错误,本文提出的方法(ISNN,Gated-ISNN)与融合历史信息的方法(CNN+LSTM)则能正确检测。

样例2中,多轮对话过程中出现了意图转换,前文语义与最后一轮的语义不相关,简单地融合上下文信息反而会干扰最后一轮的意图检测,因此只有Gated-ISNN和Attention-RNN正确检测此样例。

本节验证了本文提出的方法的优越性,能够在利用上下文信息补充语义缺失的同时控制信息共享的规模以获得优良的意图检测性能。

结束语 针对口语文本的意图检测任务面临的各项挑战,本文提出了一种基于门控机制的信息共享网络,该网络首先在文本嵌入模块结合字音特征来减弱语音解析错误带来的负面影响;其次,在语义表示模块提出了基于层级化注意力机制的多级语义编码模型,得到当前轮与上下文文本的深层语义表示;最后,通过基于门控机制的信息共享模块将上下文信息用于辅助当前轮文本的意图检测。

实验结果表明,本文提出的方法与现有的各项研究相比,性能取得了显著的提升,证明了基于上下文信息的信息共享网络方法能够有效地检测口语文本意图。

在未来的工作中,我们将会不断地完善现有的方法,以进一步提升口语意图检测的性能。同时,还将对本文方法中的模型做进一步的拓展与探索,尝试在网络中融合与意图检测相关的其他各项口语理解任务,以更好地应用于对话系统的研究。

参 考 文 献

[1] WANG Y, REN F J, QUAN C Q. A Summary of Research on Dialogue Management Methods in Spoken Dialogue System[J]. Computer Science, 2015, 42(6): 1-7, 27.

[2] CHEN H, LIU X, YIN D, et al. A survey on dialogue systems: Recent advances and new frontiers[J]. ACM SIGKDD Explorations Newsletter, 2017, 19(2): 25-35.

[3] HENDERSON M, THOMSON B, WILLIAMS J D. The second dialog state tracking challenge[C]// Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL). ACL. Stroudsburg, PA. 2014: 263-272.

[4] ZONG C Q, WU H, HUANG T Y, et al. Analysis of Spoken Dialog Corpus in Restricted Domain[C]// Proceedings of the 5th National Conference on Computational Languages. 1999: 115-122.

[5] SONG H Y, ZHANG W N, LIU T. DQN based Policy Learning for Open Domain Multi-turn Dialogues[J]. Journal of Chinese Information Processing, 2018, 32(7): 99-108, 136.

[6] SENEFF S. TINA: A natural language system for spoken language applications[J]. Computational Iinguistics, 1992, 18(1): 61-86.

[7] YAN P, ZHENG F, XU M. Robust parsing in spoken dialogue systems[C]// Seventh European Conference on Speech Communication and Technology. Academic. Amsterdam. 2001.

[8] HUANG Y F, ZHENG F, YAN P J, et al. The Design and Implementation of Campus Navigation System: Easy Nav[J]. Journal of Chinese Information Processing, 2001, 15(4): 36-41.

[9] DENG Y, XU B, HUANG T. Chinese spoken language understanding across domain[C]// Sixth International Conference on Spoken Language Processing. IEEE, 2000.

[10] MINKER W, BENNACEF S K, GAUVAIN J L. A stochastic

case frame approach for natural language understanding[C]// Fourth International Conference on Spoken Language Processing. IEEE，1996.

[11] HAFFNER P，TUR G，WRIGHT J H. Optimizing SVMs for complex call classification[C]// 2003 IEEE International Conference on Acoustics，Speech，and Signal Processing (ICASSP'03). IEEE，2003.

[12] FREUND Y，SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of computer and system sciences，1997，55(1)：119-139.

[13] SARIKAYA R，HINTON G E，RAMABHADRAN B. Deep belief nets for natural language call-routing [C]// Proceeding of the IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). IEEE，2011：5680-5683.

[14] TUR G，DENG L，HAKKANI-TÜR D，et al. Towards deeper understanding：Deep convex networks for semantic utterance classification [C]// Proceeding of the IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). IEEE，2012：5045-5048.

[15] HASHEMI H B，ASIAEE A，KRAFT R. Query intent detection using convolutional neural networks [C]// International Conference on Web Search and Data Mining，Workshop on Query Understanding. New York：ACM，2016.

[16] RAVURI S，STOICKE A. A comparative study of neural network models for lexical intent classification[C]// 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE，2015：368-374.

[17] LIU B，LANE I. Attention-based recurrent neural network models for joint intent detection and slot filling[J]. arXiv：1609.01454.

[18] FIRDAUS M，BHATNAGAR S，EKBAL A，et al. Intent Detection for Spoken Language Understanding Using a Deep Ensemble Model[C]// Pacific Rim International Conference on Artificial Intelligence. Cham：Springer，2018：629-642.

[19] BARAHONA L M R，GASIC M，MRKŠIĆ N，et al. Exploiting Sentence and Context Representations in Deep Neural Models for Spoken Language Understanding[J]. arXiv：1610.04120，2016.

[20] XIE Z，LING G. Dialogue Breakdown Detection using Hierarchical Bi-Directional LSTMs [C]// Proceedings of the Dialog System Technology Challenges Workshop (DSTC6). Elsevier. Amsterdam. 2017.

[21] BENGIO Y，DUCHARME R，VINCENT P，et al. A neural probabilistic language model[J]. Journal of machine learning research，2003，3(2)：1137-1155.

[22] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient estimation of word representations in vector space[J]. arXiv：1301.3781，2013.

[23] ZHANG X，ZHAO J，LECUN Y. Character-level convolutional networks for text classification [C]// Advances in Neural Information Processing Systems. New York：Curran Associates，2015：649-657.

[24] ZHANG X，LECUN Y. Which Encoding is the Best for Text Classification in Chinese，English，Japanese and Korean? [J]. arXiv：1708.02657，2017.

[25] SORDONI A，BENGIO Y，VAHABI H，et al. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion [C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. New York：ACM，2015：553-562.

[26] HOCHREITER S，SCHMIDHUBER J. Long short-term memory[J]. Neural Computation，1997，9(8)：1735-1780.

[27] BAHDANAU D，CHO K，BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv：1409.0473，2014.

[28] CARUNA R. Multitask learning：A knowledge based source of inductive bias[C]// Machine Learning：Proceedings of the Tenth International Conference. New York：ACM，1993：41-48.

[29] LIU P，QIU X，HUANG X. Recurrent neural network for text classification with multi-task learning[J]. arXiv：1605.05101，2016.

[30] SØGAARD A，GOLDBERG Y. Deep multi-task learning with low level tasks supervised at lower layers[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. ACL：Stroudsburg. 2016：231-235.

[31] KINGMA D P，BA J. Adam：A method for stochastic optimization[J]. arXiv：1412.6980，2014.

[32] WILLIAMS J D. Web-style ranking and SLU combination for dialog state tracking [C]// Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL). Stroudsburg：ACL，2014：282-291.

[33] WANG Y，SHEN Y，JIN H. A Bi-model based RNN Semantic Frame Parsing Model for Intent Detection and Slot Filling[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics Stroudsburg：ACP，2018：309-314.

XU Yang，born in 1994，postgraduate，is member of China Computer Federation (CCF). His main research interests include natural language processing and Dialogue system.

LI Shou-shan，born in 1980，professor，is member of China Computer Federation (CCF). His main research interests include natural language processing，dialogue system and emotion analysis.