

基于BERT的意图分类与槽填充联合方法

覃俊^{1,2}, 马天宇^{1,2}, 刘晶^{*1,2}, 帖军^{1,2}, 后琦^{1,2}

1.中南民族大学 计算机科学学院, 湖北省 武汉市 430074

2.湖北省制造企业智能管理工程技术研究中心, 湖北省 武汉市 430074

{498011695,425290717}@qq.com,{jingl,juntie}@mail.scuec.edu.cn,928983003@qq.com

摘要

口语理解是自然语言处理的一个重要内容,意图分类和槽填充是口语理解的两个基本子任务。最近的研究表明,共同学习这两项任务可以起到相互促进的作用。本文提出了一个基于BERT的意图分类联合模型,通过一个关联网络使得两个任务建立直接联系,共享信息,以此来提升任务效果。模型引入BERT来增强词向量的语义表示,有效解决了目前联合模型由于训练数据规模较小导致的泛化能力较差的问题。实验结果表明,该模型能有效提升意图分类和槽填充的性能。

关键词: 意图分类; 槽填充; BERT; 关联网络

Joint Method of Intention Classification and Slot Filling Based on BERT

Jun Qin^{1,2}, Tianyu Ma^{1,2}, Jing Liu^{*1,2}, Jun Tie^{1,2}, Qi Hou^{1,2}

1.School of Computer Science, South-Central University for Nationalities, Wuhan 430074

2.Hubei Provincial Engineering Research Center for Intelligent Management of Manufacturing Enterprises, Wuhan 430074

{498011695,425290717}@qq.com,{jingl,juntie}@mail.scuec.edu.cn,928983003@qq.com

Abstract

Spoken language understanding is an important part of natural language processing. Intent classification and slot filling are two basic sub-tasks of spoken language understanding. Recent research has shown that learning these two tasks together can be mutually reinforcing. In this paper, a joint model of intention classification based on BERT is proposed. Through an association network, the two tasks establish direct contact and share information, so as to improve the effectiveness of the task. Bert is introduced into the model to enhance the semantic representation of word vectors, which effectively solves the problem that the current joint model has poor generalization ability due to the small size of training data. Experimental results show that the proposed model can significantly improve the accuracy of intention classification and slot filling F1 value on ATIS and Snips data sets.

Keywords: Intent Classification, Slot Cilling, BERT, Association Network

*通讯作者

©2021 中国计算语言学大会

根据《Creative Commons Attribution 4.0 International License》许可出版

1 引言

随着人工智能时代的发展,越来越多的虚拟助手应用于我们的生活之中,如百度的小度、苹果的Siri、小米的小爱同学和微软的Cortana,人机对话系统已经成为人工智能和人机交互中的核心技术。**口语理解(SLU)**是让计算机能够正确理解人类的自然语言,进而为用户反馈正确信息的关键环节。口语理解通常涉及意图分类和槽填充两个子任务,在人机对话系统中起着重要的作用。意图分类侧重于从话语中自动检测用户的意图,可以将其视为预测用户意图的句子分类问题。早期的意图分类主要依靠规则进行(Ramanand et al., 2010),随后,机器学习技术被引入意图分类任务(Thomson, 2013), (Evgeniou et al., 1999))。随着深度学习的方法越来越多样和成熟,各种深度学习的网络(Suman et al., 2015) (Rahul and Salem, 2017))被应用于在意图识别任务中。与传统的基于机器学习的方法相比,深度学习体系结构可以捕获远程依赖关系并自动捕获单词之间的关联,在识别性能上有了很大的提升。槽填充是指从输入文本中提取语义成分,并在语义框架中为一组预定义的槽来填充值,为话语中的每个单词分配语义标签,该任务通常被认为是序列标注问题。通常采用BIO标注的方式来进行标注,即将每个元素标注为“B-X”、“I-X”或者“O”。其中,“B-X”表示此元素所在的片段属于X类型并且此元素在此片段的开头,“I-X”表示此元素所在的片段属于X类型并且此元素在此片段的中间位置,“O”表示不属于任何类型。如表1所示,从ATIS语料库中抽取的句子‘what time zone is Denver in’,从表可以看出,句子中的每一个单词对应一个槽标签,无关单词标注为0,不属于任何类型,而目的城市单词“denver”作为片段起始被标注为“B-city_name”,并且每一个句子会为其分类一个意图。

句子	what	time	zone	is	denver	in
槽标签	0	0	0	0	B-city_name	0
意图	atis_city					

Table 1: 槽注释示例

在早期的研究中,两项任务通常是分开进行的处理的,Liu等人(2015)引入的RNN-LSTM模型,只应用了一个联合损失函数来隐式地将两个任务联系起来,其中意图和槽之间没有建立明确的关系,而且这些方法都是从小规模的人类标记训练数据中提取的,因而泛化能力较差。Lane等人(2016)提出了一个基于注意的RNN模型,应用了一个联合损失函数来连接这两个任务;Zhang等人(2018)提出了一种基于胶囊网络的神经网络模型,该模型通过动态路由协议模式来完成槽填充和意图分类。本文提出了一种基于BERT的意图分类和槽填充联合学习模型ANBC(Associated Network on BERT with CRF),利用意图和槽之间的关联性为意图分类和槽填充任务建立一个关联网络进行联合训练,在槽填充任务中根据关系因子结合意图信息来生成槽信息,在意图分类任务中根据关系向量结合槽信息来生成意图信息,以此来提升任务效果。

我们的贡献:

- 1) 提出了一种基于预训练语言模型BERT的联合学习模型ANBC,构建一个双向关联的网络框架来对意图分类和槽填充建立直接关联,以增强两个任务之间的相互促进作用。
- 2) 由于BERT不能很好地学习逻辑依赖关系,提出了在BERT之上增加了CRF方法来增强槽填充任务中槽标签预测序列的准确性。
- 3) 我们在开放的公共数据集ATIS和Snips上大量的实验证明了我们模型的有效性。

2 相关工作

意图分类传统的方法是使用n-gram作为例如位置和日期等通用的实体的特征(Zhang and Wang, 2016),但这种方法由于输入空间维数的限制,无法达到一个明显的效果。随后,卷积神经网络(CNNs)和长短时记忆(LSTM)网络也被广泛应用于意图分类任务中(Kim Y, 2014) (Ravuri et al., 2015))。

项目基金:湖北省技术创新专项重大项目(2019ABA101);武汉市科技计划应用基础前沿项目(2020020601012267);中南民族大学中央高校基本科研业务费专项资金资助(CZQ20012)

近年来，一些联合模型被提出和使用，候丽仙等人(2020)提出了一种基于注意力机制的神经网络模型Attention-BLSTM用于联合意图分类和槽填充。但这只是克服了传统方法带来的误差传播问题，只采用共享参数的方式来隐式地连接两个任务，并没有对他们的相关性进行建模。Goo等人(2018)提出了一个slot-gated模型，该模型将意图信息应用于槽填充任务，并获得了优异的性能。但是在意图分类任务中没有使用槽信息，没有建立双向直接连接。

Yao等人(2014)为此任务引入了LSTM体系结构，在RNN的基础上利用转移特征和CRF的序列级优化来改进RNN，从而显式地建模输出标签的依赖关系。然而这些方法生成的槽标记的数量却与话语中的单词数量不同。为了克服这个限制，另一种方法是使用两个RNN模型作为输入编码器和输出解码器 (Liu and Lane, 2016)来构建的一个编码器-解码器模型。这种方法的优点是不需要在输入语句和输出槽标记之间对齐。

最近，预训练语言模型在利用大量未标记数据学习通用语言表征方面取得了显著的效果。一些预训练语言模型，如ELMO、GPT和BERT，在解决一些自然语言处理问题上取得了巨大的成功，如语篇分析 (Radford et al., 2018)、语义相似度计算 (Peters et al., 2018)、阅读理解和问答 (Radford et al., 2018)。ELMo使用双向LSTM结构，GPT使用从左到右的transformer结构，BERT使用双向transformer结构。虽然这样的预训练模型在语言表征方面会有显著的提升，但对于两个任务间的相互促进作用并未有很好利用。Wang (2020),Hardalov (2020)等人也使用了预训练模型BERT进行意图分类和槽填充任务的联合建模，但只考虑到将意图分类的结果给到槽，指导槽的预测，未考虑到槽信息对意图分类也是具有影响的。

3 意图分类和槽填充联合模型

在本节中，将介绍一种基于BERT的意图分类和槽填充的联合任务模型ANBC，模型的体系结构如图1所示。该模型包含三个部分：BERT层、关联网络层、CRF层。首先使用BERT对输入序列进行编码，然后将得到的意图向量和槽向量输入到关联网络层进行直接联系，在关联网络层中，槽填充任务结合关系因子与意图信息来生成槽信息，意图分类任务结合关系向量与槽信息来生成意图信息，形成进一步的促进作用，最后通过softmax函数和条件随机场 (CRF) 获取最终的分类结果和标签链。

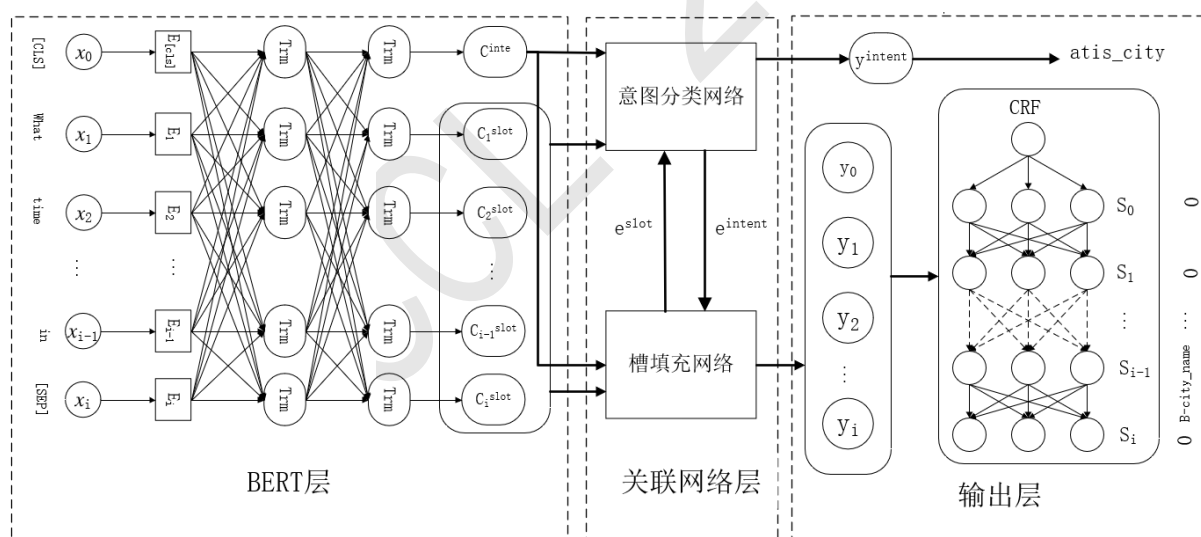


Figure 1: 体系结构

3.1 BERT层

BERT (Bidirectional Transformer Encoder) 即双向Transformer编码器，由多个双向Transformer编码器组成 (Vaswani A et al., 2017)，每一层包含一个多头自注意力子层，并且带有一个线性变换和残差连接子层。模型的输入为三个部分的总和，基于WordPiece的字向量 (Wu Y et al., 2016)、文本向量和位置向量。在大规模无标记文本上，预训练的BERT模型提供了一个强大的上下文相关的句子表示，可用于各种目标任务，对于文本分类任务，BERT模

型在文本前插入一个[CLS]符号，并将该符号对应的输出向量作为整篇文本的语义表示，还对输入的两句话用一个[SEP]符号作分割，并分别对两句话附加两个不同的文本向量以作区分。对于序列标注任务，BERT模型利用文本中每个字对应的输出向量对该字进行标注。通过BERT的多头自注意力机制，可以获得文本中每个字在不同语义空间下的增强语义向量，并将每个字的多个增强语义向量进行线性组合，从而获得一个最终的与原始字向量长度相同的增强语义向量。

在本文中，给定一个输入标记序列 $X = (x_1, \dots, x_i)$ ，则BERT的隐藏状态输出序列为 $H = (h_1, \dots, h_i)$ 。槽上下文向量 c^{slot} 由隐藏状态的加权和计算得出：

$$c^{slot} = c^{intent} = \sum_{i=1}^T \alpha_{i,j} h_i \quad (1)$$

其中权重 α 计算方式如下:

$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^T \exp(e_{i,k})} \quad (2)$$

$$e_{i,j} = g(s_{i-1}, h_k) \quad (3)$$

这里的 α 用来评估当前预测词与输入的每一个词的相关度。 g 代表一个前反馈神经网络,网络的输入就是上一个输出序列隐状态 s_{i-1} 和输入序列隐状态 h_j 。意图上下文向量 c^{intent} 的计算方法与 c^{slot} 相同,不同的是,整个句子只生成一个意图标签。最后将槽上下文向量和意图上下文向量输入到关联网络层。

3.2 关联网络层

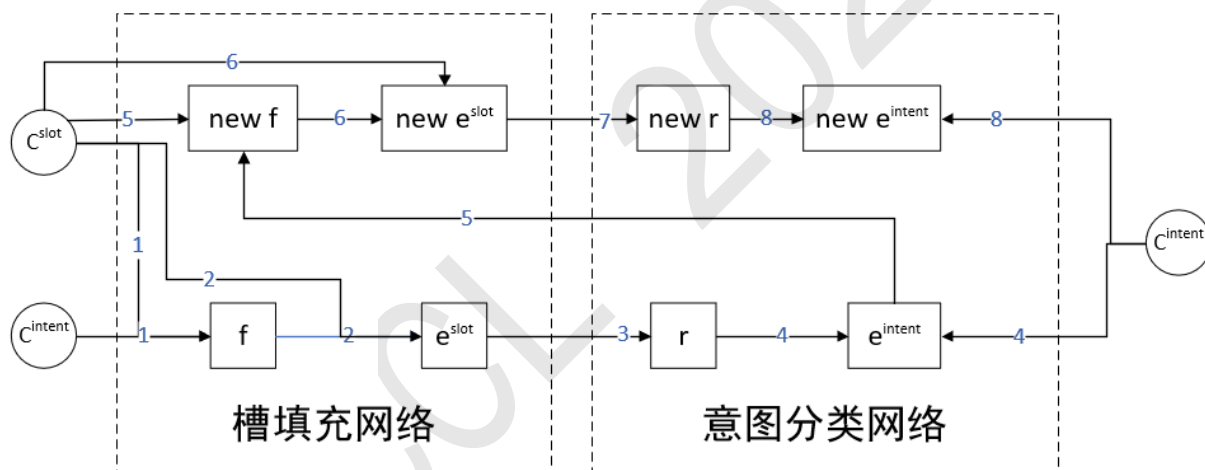


Figure 2: 关联网络示意图

关联网层由意图分类网络和槽填充网络组成如图2所示，槽填充网络中，通过一个关系因子 f 结合槽向量和意图向量生成关联槽向量输入到意图分类网络中。在意图分类网络中根据关联槽向量生成第三方关系向量 r ，再由关系向量 r 结合意图向量 c^{intent} 生成关联意图向量输回槽填充网络，以此形成一次关联迭代过程，流程如图3所示。

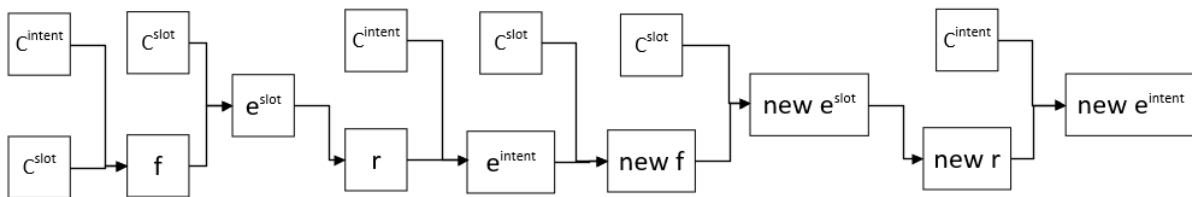


Figure 3: 迭代流程图

在槽填充网络中, 通过意图和槽向量来计算关系因子 f , 关系因子 f 通过公式4得出:

$$f = \sum V * \tanh(c_i^{slot} + W * c^{intent}) \quad (4)$$

$$r = \sum_{i=1}^T \alpha_i \cdot e_i^{slot} \quad (5)$$

$$e^{intent} = r + c^{intent} \quad (6)$$

通过关系因子 f 结合槽向量和意图向量根据公式(5)计算得出关联槽向量 e^{slot} , 然后将关联槽向量输入到意图分类网络中。

在意图分类网络中, 槽向量通过公式(6)生成关系向量 r , 并结合关系向量和意图向量, 通过公式(7)生成关联意图向量 e^{intent} , 再输入到槽填充网络中, 此时槽填充网络在接受到关联意图向量时会更新关系因子, 关系因子根据关联意图向量结合权重矩阵 W , 计算方法变为:

$$f = \sum V * \tanh(c_i^{slot} + W * e^{intent}) \quad (7)$$

其他参数的计算方法不变。

在关系因子更新之后, 槽填充网络会继续产生一个新的关联槽向量, 意图分类网络在接受到新的关联槽向量之后会对应产生一个新的关联意图向量。至此, 一次完整的关联迭代过程完成。

在关联迭代之后, 最终的意图分类和槽填充任务由关联意图向量 e^{intent} 和关联槽向量 e^{slot} 分别完成。使用关联意图向量 e^{intent} 和BERT整句文本的输出 h_T 进行最终意图分类:

$$y^{intent} = \text{softmax}(W_{h_y}^{intent} \text{concat}(h_T, e^{intent})) \quad (8)$$

对于槽填充任务, 每个字对应的输出 h_i 及其对应的关联槽向量 e_i^{slot} 进行第 i 个槽标签预测:

$$y_i^{slot} = \text{softmax}(W_{h_y}^{slot} \text{concat}(h_i, e_i^{slot})) \quad (9)$$

3.3 CRF层

槽填充本质上来讲是一个序列标记任务。对于序列标记任务, 就要考虑到邻域中标记之间的相关性。并且在很多情况下, 相邻标记的相关性是非常强的。人类的语言是事先规定了语法, 然后人类再根据语法来进行语言表达, 所以, 语法特征非常的重要, 例如, 介词后面跟名词概率很高, 之后再跟介词概率就很小。条件随机场 (CRF) 中的状态转移函数就可以很好的学习到这一特征。Zhou和Xu (2015)通过为BiLSTM编码器添加CRF层改进了语义角色标记, 表明了CRF的结构化预测模型可以改善槽填充性能, 能考虑标签序列的全局信息, 更好地对标签进行预测。核心内容如下:

给定一个句子 s , 通过把句子中所有单词的特征加权求和, 得到句子 s 的标签 l 的分数:

$$\text{score}(l|s) = \sum_{j=1}^m \sum_{i=1}^n \lambda_j f_j(s, i, l_j, l_{i-1}) \quad (10)$$

其中, i 用来表示句子 s 中第 i 个单词; l_i 表示要评分的标注序列给第 i 个单词标注的词性; l_{i-1} 表示要评分的标注序列给第 $i-1$ 个单词标注的词性。

f_j 为特征函数, 是一种序列与标签位置关系的规定, 序列标注任务是通过特征函数来定义: 对于一个序列来说, 如果前一个词的标签为动词, 那么后面一个词的标签就是名词, 也就是说定义了一个规则: 动词后面跟名词, 可以表示为名词——>动词, 而这个规则作为一个函数, 要有输出, 最简单的就是满足规则输出1, 否则输出0, 特征函数的数量根据标签种类的不同而有所变化。

λ_j 表示为不同特征函数对应的权重。不同特征函数对于模型价值是不一样的。可能在训练完后, 某个特征函数只有一次输出为1, 而另一个特征函数多次输出为1, 那么前一个特征函数对模型的价值和后一个特征函数对模型的价值就是不同的, 我们有理由根据特征函数在训练过

程中的表现对其重要性进行评测，从而最终得到每个特征函数的合适的权重，然后与特征函数相乘。可以看出，在训练过程中权重是在不断变化的，而权重也就是CRF作为一种传统机器学习算法所需要训练的参数。

第一个积分表示每个单词在所有特征函数的求和，共有 m 个特征函数；第二个积分表示对句子的每个单词求和，共有 n 个单词。定义好一组特征函数后，我们要给每个特征函数 f_j 赋予一个权重 λ_j 。现在，只要有一个句子 s ，有一个标注序列 l ，我们就可以利用前面定义的特征函数集来对 l 评分。

对这个分数进行指数化和标准化，我们就可以得到标注序列 l 的概率值 $p(l|s)$ ：

$$p(l|s) = \frac{\exp[\text{score}(l|s)]}{\sum_l \exp(\text{score}(l|s))} \quad (11)$$

由上式可知，给定一个句子 s ，标注序列 l 的分数越高，则相应的条件概率也越大，选择条件概率最大的标注序列 l 为句子 s 的标注序列。

4 实验与评估

4.1 数据集

本文中在两个广泛使用的数据集上进行了实验：航空公司旅行信息系统数据集（ATIS）和Snips。ATIS包含预订航班的人员的录音文本，训练集包含4,978语音文本，测试集包含893语音文本。训练集中有120种槽标签和21种意图类型。Snips数据集是Snips语音助手收集的对话数据集。训练集包含13,784语音文本，测试集包含700语音文本。有72个槽标签和7种意图类型。

ATIS和Snips都是英文数据集，其中ATIS是关于航班信息的数据集。而Snips比ATIS更复杂，具有更多类型的意图和更丰富的数据，包括天气，预订餐厅和播放音乐。

Dataset	ATIS	Snips
训练集	4478	13085
测试集	893	700
验证集	500	700
槽标签	120	72
意图类型	21	7

Table 2: ATIS片段和Snips数据集的统计信息

Epochs	Intent	Slot
50	99.0	98.01
40	99.14	97.95
30	98.78	97.09
20	98.3	96.4
10	97.53	97.09
5	98	96.64

Table 3: Snips数据集上的Epochs对比

4.2 评估指标及实验设置

我们使用三个评估指标来评估我们的模型。对于槽填充任务，应用F1分数。对于意图分类任务，使用准确度intent accuracy。此外，句子级的语义框架准确度sentence accuracy（句子准确度）用来表示两个任务的总体表现，即句子在整个语料库中的位置和意图都被正确预测的句子所占的比例。

本文实验中使用的是Google发布的BERT Base版本，该模型采用12层的Transformer，隐藏大小为768，自注意力的multi-head为12，模型的所有参数为110M。在训练过程中，采用

了adam优化 (Kingma and Jimmy, 2014)。初始学习率设置为0.001, 衰减指数为0.9。Epochs如表3所示, 选取效果最好的Epochs, 即40。

4.3 评估指标及实验设置

在本小节中, 我们将ANBC与现有的基线进行比较, 包括:

(1) CAPSULE-NLU (Zhang et al., 2018):该模型采用基于胶囊的神经网络模型, 利用语义层次进行建模, 并通过动态路由协议模式实现槽填充和意图分类。

(2) Attention BiRNN (Liu,B.,Lane,I., 2016):该模型引入了一种基于RNN的联合槽填充和意图分类的编解码模型,将槽填充任务视为生成任务, 应用序列到序列RNN来标记输入。此外, 使用编码器隐藏状态上的注意力加权来分类意图。

(3) Slot-Gated Full Atten (Coucke et al., 2018):该模型引入了槽选通机制作为一种特殊的选通函数, 用于集成全局意图信息以改进槽预测, 使用意图上下文向量用于意图分类。

(4) Joint Seq (Celikyilmaz et al., 2016):使用递归神经网络(RNN)获取槽填充序列中每个标记的隐藏状态, 并使用最后一个状态预测意图。

(5) Joint BERT (Chen et al., 2019):采用标准的BERT分类和序列标注分类来共同建模槽和意图。

(6) SASGBC (Wang et al., 2020):采用预训练BERT模型来解决泛化问题, 并通过选通机制来利用槽和意图之间的语义关联。

(7) BERT-Cap (Liu et al., 2020): 一种基于预训练BERT和胶囊网络的具有焦点损失的BERT-Cap混合模型。

Models	ATIS数据集			Snips数据集		
	Slot(F1)	Intent(Acc)	Sentence	Slot(F1)	Intent(Acc)	Sentence
Joint seq	94.3	92.6	80.7	87.3	96.9	73.2
Joint BERT	96.1	97.5	88.2	97.0	98.6	92.8
SASGBC	96.69	98.21	91.6	96.43	98.86	92.57
BERT-Cap	94.3	92.6	80.7	87.3	96.9	73.2
CAPSULE-NLU	95.2	95.0	83.4	91.8	97.3	80.9
Attention BiRNN	94.2	91.1	78.9	87.8	96.7	74.1
Slot-Gated Full Atten	94.8	93.6	82.2	88.8	97.0	75.5
ANBC	With CRF	97.89	98.1	91.9	98.04	99.14
	Without CRF	97.01	98.7	92.1	97.95	99.0

Table 4: ATIS和Snips数据集上不同联合模型训练结果

从表4中可以看出, 本文提出的模型ANBC很大程度上优于所有的基线模型, 该模型在槽填充 (F1)、意图分类 (Acc) 和句子准确性 (sentence) 三个方面取得了比其他模型更好的效果。在ATIS数据集上, 模型ANBC在槽填充F1得分、意图分类准确率、句子级语义框架准确率三个评价指标上的得分分别为97.89%、98.1%和92.9%的, 比当前效果最好的基线模型中对应的三个评价指标分别提高了1.2%、0.49%和1.3%。在Snips数据集上, 模型ANBC在槽填充F1得分、意图分类准确率、句子级语义框架准确率三个评价指标上的得分分别为98.04%、99.14%和91.71%, 其中槽填充F1得分和意图分类准确率比当前效果最好的基线模型得分提高了1.04%和0.28%。实验结果表明, 模型ANBC可以通过在BERT层上建立关联机制来提高SLU任务性能, 并且从表中可以看出, CRF层对模型性能有积极影响。这是因为CRF层可以在句子级别上获得最大可能的标签序列。然而, CRF层主要关注序列标记问题, 因此, 槽填充任务的改进明显超过了意图分类任务。

5 总结

在本文中, 我们提出了一种基于BERT的意图分类和槽填充联合模型ANBC, 该模型引入了预训练语言模型来更好地获取语义信息, 有效解决了长距离上下文信息依赖以及传统模型泛化能力差的问题, 并且该模型利用意图和槽之间的关联性, 为意图分类和槽填充任务建立了关

联机制，使两个任务的结果共同得到了提升。在ATIS和Snips数据集上的实验结果表明了我们的模型的有效性，实现了意图分类精度、槽填充F1精度的显著提高，可以推广到不同的数据集。在未来的工作中，我们计划在其他大规模和更复杂的自然语言理解数据集上评估所提出的方法。并且我们计划在我们提出的模型的基础上引入强化学习，尝试利用强化学习中的奖励机制来提高模型的性能。

参考文献

- Ramanand J , Bhavsar K , Pedanekar N. 2010. *Wishful thinking : finding suggestions and 'buy' wishes from product reviews*. Association for Computational Linguistics.
- B.Thomson. 2013. *Statistical methods for spoken dialogue management*. Cambridge University Press, Cambridge, UK.
- Evgeniou, Theodoros, and Massimiliano Pontil. 1999. *Support vector machines: Theory and applications*. Advanced Course on Artificial Intelligence. Springer, Berlin, Heidelberg.
- Ravuri, Suman, and Andreas Stolcke. 2015. *Recurrent neural network and LSTM models for lexical utterance classification*. Sixteenth Annual Conference of the International Speech Communication Association.
- Dey, Rahul, and Fathi M. Salem. 2017. *Gate-variants of gated recurrent unit (GRU) neural networks*. 2017 IEEE 60th international midwest symposium on circuits and systems.
- Liu, Bing, and Ian Lane. 2015. *Recurrent neural network structured output prediction for spoken language understanding*. Proc. NIPS Workshop on Machine Learning for Spoken Language Understanding and Interactions.
- Hakkani-Tür D, Tür G, Celikyilmaz A, et al. 2016. *Multi-domain joint semantic frame parsing using bi-directional rnn-lstm*. Interspeech. 2016: 715-719.
- Zhang X, Wang H. 2016. *A joint model of intent determination and slot filling for spoken language understanding*. /IJCAI. 2016, 16: 2993-2999.
- Hardalov M, Koychev I, Nakov P. 2020. *Enriched Pre-trained Transformers for Joint Slot Filling and Intent Detection*. arXiv preprint arXiv:2004.14848.
- Kim Y. 2016. 2014r08r25).<https://arXiv.org/abs/1408.5882>. *Convolutional Neural Networks for Sentence Classification*.
- Ravuri, Suman, and Andreas Stolcke. 2015. *Recurrent neural network and LSTM models for lexical utterance classification*. Sixteenth Annual Conference of the International Speech Communication Association.
- 侯丽仙,李艳玲,林民, & 李成城. 2020. 融合多约束条件的意图和语义槽填充联合识别. 计算机科学与探索, 2020(9):1545-1553.
- Goo C W, Gao G, Hsu Y K, et al. 2018. *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, Volume 2 (Short Papers).
- Yao K, Peng B, Zhang Y, et al. 2014. *Spoken language understanding using long short-term memory neural networks*. 2014 IEEE Spoken Language Technology Workshop.
- Liu, Bing, and Ian Lane. 2016. *Attention-based recurrent neural network models for joint intent detection and slot filling*. arXiv preprint arXiv:1609.01454 (2016).
- Radford A, Narasimhan K, Salimans T, et al. 2018. *Improving language understanding by generative pre-training*.
- Peters M E, Neumann M, Iyyer M, et al. 2018. *Deep contextualized word representations*. arXiv preprint arXiv:1802.05365.
- Radford A, Wu J, Child R, et al. 2019. *Language models are unsupervised multitask learners*. OpenAI blog, 2019, 1(8):9.

- Vaswani A, Shazeer N, Parmar N, et al. 2017. *Attention is all you need*. arXiv preprint arXiv:1706.03762, 2017.
- Wu Y, Schuster M, Chen Z, et al. 2016. *Attention is all you need* Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144, 2016.
- Jie Zhou and Wei Xu. 2015. *Attention is all you need* Google's neural machine translation system: Bridging the gap between human and machine translation. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).
- Kingma, Diederik P., and Jimmy Ba. 2014. *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980.
- Zhang, C., Li, Y., Du, N., Fan, W., & Yu, P. S. 2018. *Joint slot filling and intent detection via capsule neural networks*. arXiv preprint arXiv:1812.09471.
- Liu, B., Lane, I. 2016. *Joint online spoken language understanding and language modeling with recurrent neural networks*. arXiv preprint arXiv:1609.01462(2016).
- Coucke A, Saade A, Ball A, et al. 2018. *Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces*. arXiv preprint arXiv:1805.10190.
- Hakkani-Tür D, Tür G, Celikyilmaz A, et al. 2016. *Multi-domain joint semantic frame parsing using bi-directional rnn-lstm*. Interspeech. 2016: 715-719.
- Liu, Bing, and Ian Lane. 2016. *Attention-based recurrent neural network models for joint intent detection and slot filling*. arXiv preprint arXiv:1609.01454 (2016).
- Chen Q, Zhuo Z, Wang W. 2019. *Bert for joint intent classification and slot filling*. arXiv preprint arXiv:1902.10909, 2019.
- Wang C, Huang Z, Hu M. 2020. *SASGBC: Improving Sequence Labeling Performance for Joint Learning of Slot Filling and Intent Detection*. Proceedings of 2020 the 6th International Conference on Computing and Data Engineering.
- Liu H, Liu Y, Wong L P, et al. 2020. *A Hybrid Neural Network BERT-Cap Based on Pre-Trained Language Model and Capsule Network for User Intent Classification*. Complexity 2020.