专利申请著录项目表

接案号：TF221113410

(与：　　同日申报)　　(与：　　一案两报)　　投诉二维码

客户联系人	姓名：李兒（技术） 姓名：李鹏华	电话：19823660105 电话：13648401358	电子邮箱：limao11280@163.com 电子邮箱：lipenghua88@163.com	
专利代理人	姓名：徐望　电话：023-86898822-805；13594399963		电子邮箱：thy30@thycq.com	
特别提醒：尊敬的客户，上述代理人的联系电话和邮箱为我司指定的工作电话和工作邮箱，为了保证您的技术方案的保密性和后续工作衔接的顺畅，请您务必采用上述联系方式与我方工作人员进行联系。				
①专利名称	一种基于全局-局部对比学习的跨语言自然语言理解方法			
②专利类型	发明☒　实用新型☐　外观设计☐　一案两报☐		是否费减：　是☒　否☐	
③所有发明人	李鹏华　黄子恒　张奕辉　谢潇　刘学超　唐培渊			
④第一发明人国籍	中国		身份证号	500242198412230013
⑤申请人	申请人(1)	姓名或名称	重庆邮电大学	
		组织机构代码/身份证号	125000004504018996	
		邮政编码　400065	详细地址	重庆市南岸区黄桷垭崇文路 2 号
	申请人(2)	姓名或名称		
		组织机构代码/身份证号		
		邮政编码	详细地址	
	申请人(3)	姓名或名称		
		组织机构代码/身份证号		
		邮政编码	详细地址	
⑥提前公布	☒请求提前公布该专利申请（只适用于发明专利申请）			
⑦实质审查	☒在提交专利申请的同时提交实质审查请求（只适用于发明专利申请）			

特殊专利申请信息，涉及该项内容时填写

⑧分案申请	原申请号：　　针对的分案申请号：　　原申请日：　　年　月　日
⑨生物材料样品	保藏单位：　　地址：
	保藏日期：　　年　月　日　　保藏编号：　　分类命名：
	☐在提交专利申请的同时提交生物材料样品保藏及存活证明
⑩序列表	☐本专利申请涉及核苷酸或氨基酸序列表
⑪要求优先权声明	原受理机构名称：　　在先申请日：　　年　月　日　　在先申请号：
	原受理机构名称：　　在先申请日：　　年　月　日　　在先申请号：
备注	指定说明书附图中的图___1___为摘要附图

北京同恒源知识产权代理有限公司

一、本表由代理人预先填写，请联系人仔细核对信息是否正确，若有错误或缺漏请修改或补充。

二、本表第③栏，发明人是指对发明创造的实质性特点作出创造性贡献的人。发明人应当是个人（自然人）。发明人有两个发明人可以请求国家知识产权局不公布其姓名，若请求不公布姓名，应当在此栏所填写的相应发明人后面注明“（不公布姓名）”。

三、本表第④栏，第一发明人是中国港、澳、台地区的，其国籍应填写为“中国”。第一发明人为中国内地居民的，还应当同时填写居民身份证件号码（或军官证号码）。

四、本表第⑥栏，提前公布是指在发明专利申请初步审查合格后立即进入公布准备。如果不请求提前公布，则该发明专利申请将在自申请日起满十八个月时公布。由于发明专利申请必须在公布以后才能进入实质审查程序，为了加快申请的审查进程，在申请人无特别要求的情况下，本公司默认勾选此栏。如申请人不要求提前公布，请去除勾选并及时通知代理人。

五、本表第⑦栏，实质审查是指审查员对发明专利申请是否符合授权条件（包括新颖性、创造性、实用性、公开充分、单一性问题等）进行审查。申请人可以在自申请日（有优先权的，指优先权日）起三年内提出实质审查请求来启动实质审查程序。如果在提交专利申请的同时提交实质审查请求，则该发明专利申请在公布后立即进入实质审查阶段。为了加快发明专利申请的审查进程，在申请人无特别要求的情况下，本公司默认勾选此栏。如申请人不想在提交专利申请的同时提交实质审查请求，请去除勾选并及时通知代理人。

六、本表第⑨栏，发明专利申请涉及公众不能得到的生物材料的，应当填写此栏，并自申请日起四个月内提交由保藏单位出具的该生物材料样品的保藏及存活证明。

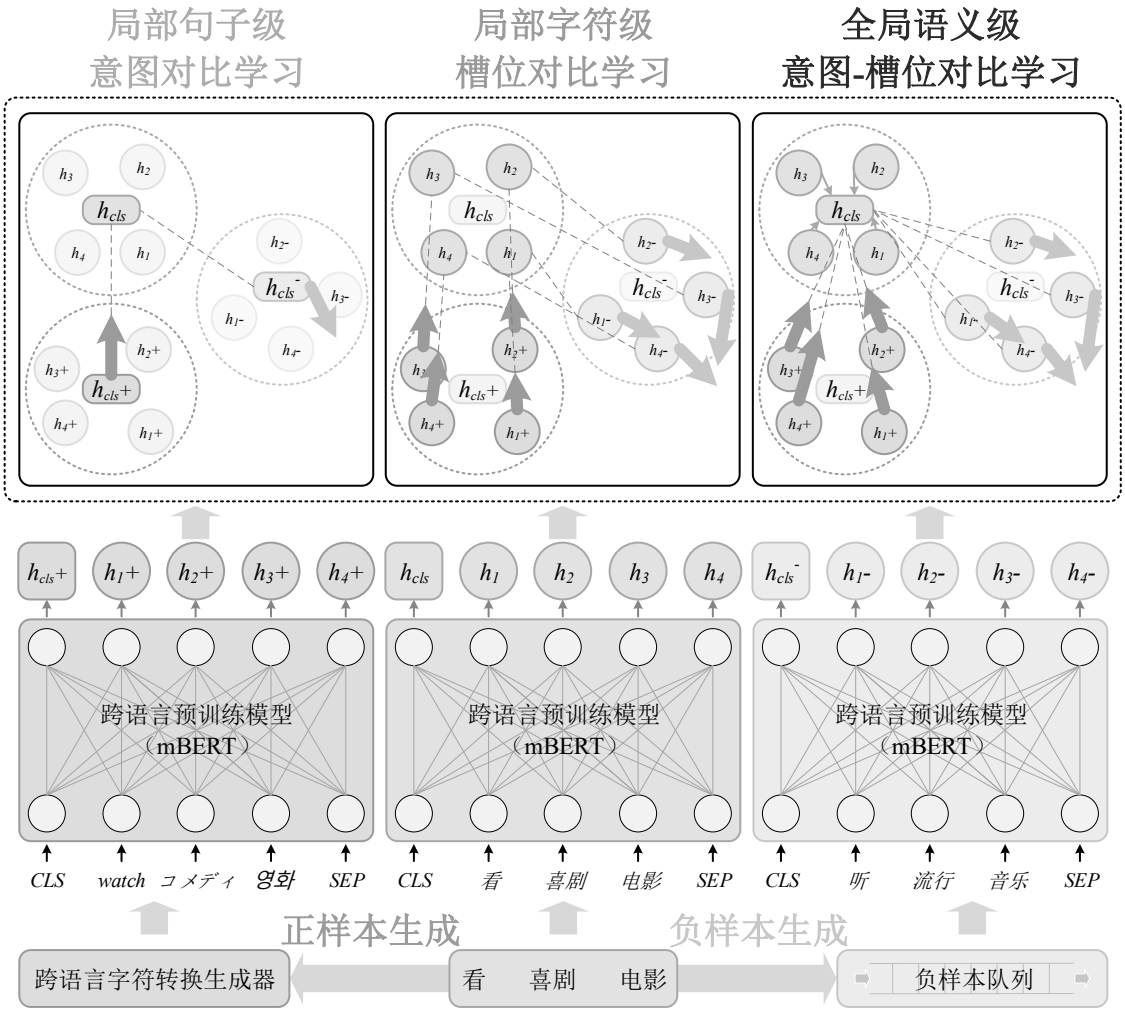
七、本表第⑩栏，发明专利申请涉及核苷酸或氨基酸序列表的，应当填写此栏，并在提交专利申请文件的同时提交核苷酸或氨基酸序列表的计算机可读形式副本。

八、本表第⑪栏，申请人要求外国或者本国优先权的，应当填写此栏。

说 明 书 摘 要

本发明涉及一种基于全局-局部对比学习的跨语言自然语言理解方法，属于自然语言处理技术领域。该方法针对自然语言理解模型的高性能跨语言迁移需求，研究基于全局-局部对比学习网络的跨语言自然语言理解方法，主要包括三个模块：局部句子级意图对比学习模块，针对意图检测任务实现跨语言句子表示对齐；局部字符级槽位对比学习模块，针对槽位填充任务实现跨语言字符表示对齐；语义级全局意图-槽位对比学习模块，实现意图和槽位间的表示对齐。本发明能够学习不同层级的细粒度对齐信息，挖掘出丰富的语义特征，缩小原始语言与目标语言之间的预测差异。

摘 要 附 图


局部句子级
意图对比学习
局部字符级
槽位对比学习
全局语义级
意图-槽位对比学习
跨语言预训练模型
（mBERT）
跨语言预训练模型
（mBERT）
跨语言预训练模型
（mBERT）
CLS watch コメディ 영화 SEP
CLS 看 喜剧 电影 SEP
CLS 听 流行 音乐 SEP
正样本生成
负样本生成
跨语言字符转换生成器
看 喜剧 电影
负样本队列

权 利 要 求 书

1、一种基于全局-局部对比学习的跨语言自然语言理解方法，其特征在于：该方法包括以下步骤：

S1、生成原始话语序列，根据跨语言字典将原始话语序列翻译为正样本，将正样本输入到跨语言预训练模型中得到对应的编码表示；

S2、；根据经过编码的原始话语序列、正样本以及前一时刻的负样本生成负样本队列，将负样本队列输入到跨语言预训练模型中得到对应的编码表示；

S3、通过建立损失函数来构建局部句子级意图对比学习模块，实现跨语言句子表示对齐；

S4、通过建立损失函数来构建局部字符级槽位对比学习模块，实现跨语言字符表示对齐；

S5、通过建立损失函数来构建全局语义级意图-槽位对比学习模块，实现意图和槽位的表示对齐，完成跨语言理解。

2、根据权利要求 1 所述的跨语言自然语言理解方法，其特征在于：在步骤 S1 中，对于原始话语序列中的每个字符，在跨语言字典中随机选择相应的翻译字符进行替换以生成正样本；

将正样本输入到预训练模型中，通过其中的双向循环神经网络生成隐层状态表示 $h_i = BiLSTM(\theta^{emb}(x_i), h_{i-1}, h_{i+1})$，其中 θ^{emb} 表示向量化函数，最终可得到针对正样本的编码表示为：

$$H_+ = (h_{CLS^+}, h_{1^+}, ..., h_{n^+}, h_{SEP^+})$$

式中，h_{CLS^+}、h_{SEP^+} 分别表示正样本开始标志位的向量表示和结束标志位的向量表示，$h_{[]^+}$ 表示正样本中各字符被编码后形成的向量表示。

3、根据权利要求 1 所述的跨语言自然语言理解方法，其特征在于：在步骤 S2 中，以类似步骤 S1 的方法将负样本输入到预训练模型中得到的编码表示为：

$$H_{CLS^-} = \left(h^k_{CLS^-}\right)^{K-1}_{k=0}, H_{S^-} = \left(h^k_{S^-}\right)^{K-1}_{k=0}$$

式中，K 表示负样本队列的最大容量，$h^k_{CLS^-}$ 表示负样本队列开始标志位的向量表示，$h^k_{S^-}$ 表示负样本队列中各字符被编码后形成的向量表示。

4、根据权利要求 1 所述的跨语言自然语言理解方法，其特征在于：在步骤 S3 中，通过设计损失函数来构建局部句子级意图对比学习模块，损失函数如下式：

$$\mathcal{L}_{CL-Y} = -\ln \frac{s\left(h_{CLS}, h_{CLS^+}\right)}{s\left(h_{CLS}, h_{CLS^+}\right) + \sum_{k=0}^{K-1} s\left(h_{CLS}, h^k_{CLS^-}\right)}$$

式中，$s([],[])$ 表示点积操作，h_{CLS} 表示原始话语序列开始标志位的向量表示，h_{CLS^+} 表示

正样本中开始标志位的向量表示，$h_{CLS^-}^k$ 表示负样本队列中开始标志位的向量表示，K 表示负样本队列的最大容量。

5、根据权利要求 1 所述的跨语言自然语言理解方法，其特征在于：在步骤 S4 中，通过设计损失函数来构建局部字符级槽位对比学习模块，损失函数如下式：

$$\mathcal{L}_{CL-C}^{i} = -\frac{1}{n}\sum_{j=1}^{n}\ln\frac{s(h_i, h_{j^+})}{s(h_i, h_{j^+}) + \sum_{k=0}^{K-1} s(h_i, h_{j^-}^k)}$$

式中，$s([],[])$表示点积操作，n 表示原始话语序列的长度，h_i 表示原始话语序列中位置 i 处字符的向量表示，h_{j^+} 表示正样本中位置 j 处字符的向量表示，$h_{j^-}^k$ 表示负样本队列中位置 j 处字符的向量表示，K 表示负样本队列的最大容量。

6、根据权利要求 5 所述的跨语言自然语言理解方法，其特征在于：在局部字符级槽位对比学习模块中，最终的损失 $\mathcal{L}_{CL-C}$ 为所有字符损失函数的总和。

7、根据权利要求 1 所述的跨语言自然语言理解方法，其特征在于：在步骤 S5 中，通过设计损失函数来构建全局语义级意图-槽位对比学习模块，损失函数设计如下所述：

首先分别构建针对原始话语序列和正样本序列的损失函数：

$$\mathcal{L}_{G1} = \frac{1}{n}\sum_{j=1}^{n}\ln\frac{s\left(h_{CLS}, h_j\right)}{s\left(h_{CLS}, h_j\right) + \sum_{k=0}^{K-1} s\left(h_{CLS}, h_{j^-}^k\right)}$$

$$\mathcal{L}_{G2} = \frac{1}{n}\sum_{j=1}^{n}\ln\frac{s\left(h_{CLS}, h_{j^+}\right)}{s\left(h_{CLS}, h_{j^+}\right) + \sum_{k=0}^{K-1} s\left(h_{CLS}, h_{j^-}^k\right)}$$

式中，$\mathcal{L}_{G1}$ 表示针对原始话语序列的损失函数，$\mathcal{L}_{G2}$ 表示针对正样本的损失函数，n 表示原始话语序列的长度，h_{CLS} 表示原始话语话语序列开始标志位的向量表示，h_j 表示原始话语序列中位置 j 处字符的向量表示，$h_{j^-}^k$ 表示负样本队列中位置 j 处字符的向量表示，h_{j^+} 表示正样本中位置 j 处字符的向量表示，K 表示负样本队列的最大容量；

结合损失函数 $\mathcal{L}_{G1}$ 和 $\mathcal{L}_{G2}$，得到针对语义级对比学习的损失函数 $\mathcal{L}_G$：

$$\mathcal{L}_G = \mathcal{L}_{G1} + \mathcal{L}_{G2}$$

然后分别构建针对意图检测的损失函数 $\mathcal{L}_Y$，以及针对槽位填充的损失函数 $\mathcal{L}_C$：

$$\mathcal{L}_Y \triangleq -\sum_{i=1}^{n_Y} \hat{\kappa}_i^Y \ln(\sigma^Y)$$

$$\mathcal{L}_C \triangleq -\sum_{j=1}^{n}\sum_{i=1}^{n_C}\hat{\kappa}_j^{i,C}\ln\left(\kappa_j^{i,C}\right)$$

式中，$\sigma^Y=\mathrm{softmax}(\boldsymbol{W}^Y\boldsymbol{h}_{\mathrm{CLS}}+\boldsymbol{b}^Y)$，其中，$W^Y$和$b^Y$表示可训练的参数，$\hat{\kappa}_i^Y$表示意图标签，$\hat{\kappa}_j^{i,C}$表示针对位置为 i 的字符的槽位标签的预测值，$\kappa_j^{i,C}$表示针对位置为 i 的字符的槽位标签的实际值，n_Y表示意图标签的数量，n_C表示槽位标签的数量；

最后结合损失函数$\mathcal{L}_Y$、$\mathcal{L}_C$、$\mathcal{L}_G$、$\mathcal{L}_{CL-C}$和$\mathcal{L}_{CL-Y}$得到全局语义级意图-槽位对比学习的总体损失函数：

$$\mathcal{L}_{total}=\lambda_1\mathcal{L}_Y+\lambda_2\mathcal{L}_C+\lambda_3\mathcal{L}_{CL-Y}+\lambda_4\mathcal{L}_{CL-C}+\lambda_5\mathcal{L}_{\mathrm{G}}$$

式中，λ表示可训练的参数。

一种基于全局-局部对比学习的跨语言自然语言理解方法

技术领域

本发明属于自然语言处理技术领域，涉及一种基于全局-局部对比学习的跨语言自然语言理解方法。

背景技术

目前，语言依旧是人流交流信息的第一载体，是一种最为有效、便捷的方式，语音交互作为人机通信中最自然、直接的交互方式，具有天然的优势。作为其中的一项关键技术，自然语言理解通常包含意图检测和槽位填充两个子任务。为了使自然语言理解模型能够更好的应用于缺乏大量标记数据的低资源语言，许多研究都聚焦于使用零样本学习构建网络，零样本学习的方法可以利用高资源语言中的标记数据来训练模型，并将其转移到目标低资源语言上得以应用。

虽然零样本学习能大大减少人工标记数据的工作量，并且也在领域内实现了很好的效果，但是该方法仅依赖于共享参数，并且只能执行跨语言的隐式对齐。这种机制带来了两个问题，其一，这种隐式对齐的过程在目前看来还是一个黑箱，不仅严重影响对齐表示，而且难以分析其对齐机制；其二，许多研究工作并没有充分考虑到两个子任务的不同细粒度层级，例如：意图检测是句子级的，而槽位填充是字符级的，这会导致意图和槽位之间无法相互接收一些来自不同粒度层级的迁移信息，影响了模型的预测性能。因此，现在的工作在于弥补现有基于零样本学习的自然语言理解模型在对齐机制和子任务交互方面存在的缺陷。

发明内容

有鉴于此，本发明的目的在于提供一种基于全局-局部对比学习的跨语言自然语言理解方法，通过对比学习方法将不同语种的相似句子表示显式对齐，使用局部对比学习来学习不同层级的细粒度对齐信息，实现全局-局部信息融合，完成跨语言理解。

为达到上述目的，本发明提供如下技术方案：

一种基于全局-局部对比学习的跨语言自然语言理解方法，该方法包括以下步骤：

S1、生成原始话语序列，根据跨语言字典将原始话语序列翻译为正样本，将正样本输入到跨语言预训练模型中得到对应的编码表示；

S2、；根据经过编码的原始话语序列、正样本以及前一时刻的负样本生成负样本队列，将负样本队列输入到跨语言预训练模型中得到对应的编码表示；

S3、构建局部句子级意图对比学习模块，实现跨语言句子表示对齐；

S4、构建局部字符级槽位对比学习模块，实现跨语言字符表示对齐；

S5、构建全局语义级意图-槽位对比学习模块，实现意图和槽位的表示对齐。

进一步，在步骤 S1 中，对于原始话语序列中的每个字符，在跨语言字典中随机选择相应的翻译字符进行替换以生成正样本；

将正样本输入到预训练模型中得到的编码表示为：

$$H_{+}=(h_{CLS^{+}},h_{1^{+}},...,h_{n^{+}},h_{SEP^{+}})$$

式中，$h_{CLS^{+}}$、$h_{SEP^{+}}$分别表示正样本开始标志位的向量表示和结束标志位的向量表示，$h_{[]^{+}}$表示正样本中各字符被编码后形成的向量表示。

进一步，在步骤 S2 中，将负样本输入到预训练模型中得到的编码表示为：

$$H_{CLS^{-}}=\left(h_{CLS^{-}}^{k}\right)_{k=0}^{K-1},H_{S^{-}}=\left(h_{S^{-}}^{k}\right)_{k=0}^{K-1}$$

式中，K 表示负样本队列的最大容量，$h_{CLS^{-}}^{k}$ 表示负样本队列开始标志位的向量表示，$h_{S^{-}}^{k}$ 表示负样本队列中各字符被编码后形成的向量表示。

进一步，在步骤 S3 中，通过设计损失函数来构建局部句子级意图对比学习模块，损失函数如下式：

$$\mathcal{L}_{CL-Y}=-\ln\frac{s\left(h_{CLS},h_{CLS^{+}}\right)}{s\left(h_{CLS},h_{CLS^{+}}\right)+\sum_{k=0}^{K-1}s\left(h_{CLS},h_{CLS^{-}}^{k}\right)}$$

式中，$s([],[])$表示点积操作，h_{CLS} 表示原始话语序列开始标志位的向量表示，$h_{CLS^{-}}^{k}$ 表示负样本队列中开始标志位的向量表示，K 表示负样本队列的最大容量。

进一步，在步骤 S4 中，通过设计损失函数来构建局部字符级槽位对比学习模块，损失函数如下式：

$$\mathcal{L}_{CL-C}^{i}=-\frac{1}{n}\sum_{j=1}^{n}\ln\frac{s(h_{i},h_{j^{+}})}{s(h_{i},h_{j^{+}})+\sum_{k=0}^{K-1}s(h_{i},h_{j^{-}}^{k})}$$

式中，n 表示原始话语序列的长度，h_{i} 表示原始话语序列中位置 i 处字符的向量表示，$h_{j^{+}}$ 表示正样本中位置 j 处字符的向量表示，$h_{j^{-}}^{k}$ 表示负样本队列中位置 j 处字符的向量表示。

进一步，在局部字符级槽位对比学习模块中，最终的损失 $\mathcal{L}_{CL-C}$ 为所有字符损失函数的总和。

进一步，在步骤 S5 中，通过设计损失函数来构建全局语义级意图-槽位对比学习模块，损失函数设计如下所述：

首先分别构建针对原始话语序列和正样本序列的损失函数：

$$\mathcal{L}_{G1}=\frac{1}{n}\sum_{j=1}^{n}\ln\frac{s\left(h_{CLS},h_j\right)}{s\left(h_{CLS},h_j\right)+\sum_{k=0}^{K-1}s\left(h_{CLS},h_{j^-}^{k}\right)}$$

$$\mathcal{L}_{G2}=\frac{1}{n}\sum_{j=1}^{n}\ln\frac{s\left(h_{CLS},h_{j^+}\right)}{s\left(h_{CLS},h_{j^+}\right)+\sum_{k=0}^{K-1}s\left(h_{CLS},h_{j^-}^{k}\right)}$$

式中，$\mathcal{L}_{G1}$ 表示针对原始话语序列的损失函数，$\mathcal{L}_{G2}$ 表示针对正样本的损失函数，h_j 表示原始话语序列中位置 j 处字符的向量表示。

结合损失函数 $\mathcal{L}_{G1}$ 和 $\mathcal{L}_{G2}$，得到针对语义级对比学习的损失函数 $\mathcal{L}_G$：

$$\mathcal{L}_G=\mathcal{L}_{G1}+\mathcal{L}_{G2}$$

然后分别构建针对意图检测的损失函数 $\mathcal{L}_Y$，以及针对槽位填充的损失函数 $\mathcal{L}_C$：

$$\mathcal{L}_Y\triangleq-\sum_{i=1}^{n_Y}\hat{\kappa}_i^Y\ln(\sigma^Y)$$

$$\mathcal{L}_C\triangleq-\sum_{j=1}^{n}\sum_{i=1}^{n_C}\hat{\kappa}_j^{i,C}\ln\left(\kappa_j^{i,C}\right)$$

式中，$\sigma^Y=\mathrm{softmax}(\boldsymbol{W}^Y\boldsymbol{h}_{\mathrm{CLS}}+\boldsymbol{b}^Y)$，其中，$W^Y$ 和 b^Y 表示可训练的参数，$\hat{\kappa}_i^Y$ 表示意图标签，$\hat{\kappa}_j^{i,C}$ 表示针对位置为 i 的字符的槽位标签的预测值，$\kappa_j^{i,C}$ 表示针对位置为 i 的字符的槽位标签的实际值，n_Y 表示意图标签的数量，n_C 表示槽位标签的数量。

最后结合损失函数 $\mathcal{L}_Y$、$\mathcal{L}_C$、$\mathcal{L}_G$、$\mathcal{L}_{CL-C}$ 和 $\mathcal{L}_{CL-Y}$ 得到全局语义级意图-槽位对比学习的总体损失函数：

$$\mathcal{L}_{total}=\lambda_1\mathcal{L}_Y+\lambda_2\mathcal{L}_C+\lambda_3\mathcal{L}_{GC-Y}+\lambda_4\mathcal{L}_{GC-C}+\lambda_5\mathcal{L}_{\mathrm{G}}$$

式中，λ 表示可训练的参数，均为标量。

本发明的有益效果在于：本发明通过对比学习方法将不同语种的相似句子表示显式对齐，支持对对齐方式进行分析；使用局部对比学习模块来学习意图和槽位中不同层级的细粒度对齐信息，同时利用全局对比学习模块构建意图和槽位的交互通道以挖掘更丰富的语义特征，实现全局-局部信息融合，完成跨语言理解，缩小了原始语言和目标语言之间的预测差异。

本发明的其他优点、目标和特征在某种程度上将在随后的说明书中进行阐述，并且在某种程度上，基于对下文的考察研究对本领域技术人员而言将是显而易见的，或者可以从本发

明的实践中得到教导。本发明的目标和其他优点可以通过下面的说明书来实现和获得。

附图说明

为了使本发明的目的、技术方案和优点更加清楚，下面将结合附图对本发明作优选的详细描述，其中：

图 1 为本发明模型框架。

具体实施方式

以下通过特定的具体实例说明本发明的实施方式，本领域技术人员可由本说明书所揭露的内容轻易地了解本发明的其他优点与功效。本发明还可以通过另外不同的具体实施方式加以实施或应用，本说明书中的各项细节也可以基于不同观点与应用，在没有背离本发明的精神下进行各种修饰或改变。需要说明的是，以下实施例中所提供的图示仅以示意方式说明本发明的基本构想，在不冲突的情况下，以下实施例及实施例中的特征可以相互组合。

本发明针对自然语言理解模型的高性能跨语言迁移需求提出，主要包括三个模块：局部句子级意图对比学习模块、局部字符级槽位对比学习模块以及语义级全局意图-槽位对比学习模块，其框架模型如图 1 所示。具体的，各模块的主要作用为：局部句子级意图对比学习模块针对意图检测任务实现跨语言句子表示对齐；局部字符级槽位对比学习模块针对槽位填充任务实现跨语言字符表示对齐；全局语义级意图-槽位对比学习模块实现意图和槽位间的对齐表示。

本发明实施步骤如下：

1、生成正、负样本。对于对比学习，关键操作是针对原始（锚）话语选择适当的正、负样本对。

1）基于跨语言字典的正样本生成

与原始话语相比，正样本应该保留相同的语义，对于长度为 n 的原始话语序列：

$$u=\left([CLS],u_1,u_2,...,[SEP]\right)$$

其中，$[CLS]$和$[SEP]$都属于标志位，前者放置于整个序列的首位，后者用于分隔非连续的序列。构建跨语言字典，在字符转换生成器的作用下生成正样本u_+。具体来说，对于序列 u 中的每个u_+，在跨语言字典中随机选择相应的翻译字符进行替换以生成正样本序列，例如，对于中文里的原始话语“看喜剧电影”，生成包含英语、日语、韩语的正样本序列“watch、コメディ、영화”，可以看作是具有相同含义的跨语言视图。将u_+输入到跨语言预训练模型（mBERT）中，即可得到相应的编码表示：

$$H_{+}=(h_{CLS^{+}},h_{1^{+}},...,h_{n^{+}},h_{SEP^{+}})$$

式中，$h_{CLS^{+}}$、$h_{SEP^{+}}$分别为正样本开始标志位的向量表示和结束标志位的向量表示，$h_{[]^{+}}$为正样本中各字符被编码后形成的向量表示。

2）基于队列机制的负样本生成

考虑到传统生成负样本的方法效率不高（例如选择当前批次中的其他字符），通过维护一个负样本队列，其中包含已编码的原始话语序列 u、正样本序列u_{+}和前一时刻的负样本序列u_{-}，能够逐步重用前一批次的样本从而减少不必要的编码过程。针对$[CLS]$的负样本队列和字句表示分别为：

$$H_{CLS^{-}}=\left(h_{CLS^{-}}^{k}\right)_{k=0}^{K-1},H_{s^{-}}=\left(H_{s^{-}}^{k}\right)_{k=0}^{K-1}$$

其中 K 表示负样本队列的最大容量，$h_{CLS^{-}}^{k}$为负样本队列开始标志位的向量表示，$h_{S^{-}}^{k}$为负样本队列中各字符被编码后形成的向量表示。

2、构建局部模块。

1）局部句子级意图对比学习模块

考虑到意图检测是句子级别的分类任务，而跨语言句子表示对齐是零样本跨语言意图检测任务的目标，因此设计一种专属的损失函数，驱动训练模型将相似的句子表示对齐到跨语言的相同局部空间以进行意图检测，损失函数如下式：

$$\mathcal{L}_{CL-Y}=-\ln\frac{s\left(h_{CLS},h_{CLS^{+}}\right)}{s\left(h_{CLS},h_{CLS^{+}}\right)+\sum_{k=0}^{K-1}s\left(h_{CLS},h_{CLS^{-}}^{k}\right)}$$

式中，$s([],[])$表示点积操作，h_{CLS}为原始话语序列开始标志位的向量表示。

2）局部字符级槽位对比学习模块

考虑到槽位填充是字符级别的标注任务，因此同样设计一种专属的损失函数来驱动训练模型完成针对槽位填充的字符对齐，实现细粒度信息的跨语言迁移。对于位置为 i 字符，损失函数为：

$$\mathcal{L}_{CL-C}^{i}=-\frac{1}{n}\sum_{j=1}^{n}\ln\frac{s(h_{i},h_{j^{+}})}{s(h_{i},h_{j^{+}})+\sum_{k=0}^{K-1}s(h_{i},h_{j^{-}}^{k})}$$

式中，n 表示原始话语序列的长度，h_{i}为原始话语序列中位置 i 处字符的向量表示，$h_{j^{+}}$为正样本中位置 j 处字符的向量表示，$h_{j^{-}}^{k}$为负样本队列中位置 j 处字符的向量表示。

最终的损失$\mathcal{L}_{CL-C}$为所有字符损失函数的总和。

3、构建全局语义级意图-槽位对比学习模块

当槽位和意图同属于一个用户查询时，它们通常在语义上高度相关，所以一个话语表现出的意图和其本身所包含的槽位可以自然地构成正样本对，而其他话语句子中对应的槽位可以形成负样本对。对此进一步地设计了针对全局语义级意图-槽位对比学习的损失函数，模拟意图和槽位之间的语义交互，进一步改善它们之间的跨语言传输性能。

首先分别构建针对原始话语序列和正样本序列的损失函数：

$$\mathcal{L}_{G1}=\frac{1}{n}\sum_{j=1}^{n}\ln\frac{s\left(h_{CLS},h_j\right)}{s\left(h_{CLS},h_j\right)+\sum_{k=0}^{K-1}s\left(h_{CLS},h_{j^-}^{k}\right)}$$

$$\mathcal{L}_{G2}=\frac{1}{n}\sum_{j=1}^{n}\ln\frac{s\left(h_{CLS},h_{j^+}\right)}{s\left(h_{CLS},h_{j^+}\right)+\sum_{k=0}^{K-1}s\left(h_{CLS},h_{j^-}^{k}\right)}$$

式中，$\mathcal{L}_{G1}$表示针对原始话语序列的损失函数，$\mathcal{L}_{G2}$表示针对正样本序列的损失函数，h_j为原始话语序列中位置 j 处字符的向量表示。

结合损失函数$\mathcal{L}_{G1}$和$\mathcal{L}_{G2}$，得到针对语义级对比学习的损失函数$\mathcal{L}_G$：

$$\mathcal{L}_G=\mathcal{L}_{G1}+\mathcal{L}_{G2}$$

然后分别构建针对意图检测的损失函数$\mathcal{L}_Y$，以及针对槽位填充的损失函数$\mathcal{L}_C$：

$$\mathcal{L}_Y\triangleq-\sum_{i=1}^{n_Y}\hat{\kappa}_i^Y\ln(\sigma_i^Y)$$

$$\mathcal{L}_C\triangleq-\sum_{j=1}^{n}\sum_{i=1}^{n_C}\hat{\kappa}_j^{i,C}\ln\left(\kappa_j^{i,C}\right)$$

式中，$\sigma^Y=\mathrm{softmax}(\boldsymbol{W}^Y\boldsymbol{h}_{\mathrm{CLS}}+\boldsymbol{b}^Y)$，其中，$W^Y$和$b^Y$表示可训练的参数，$\hat{\kappa}_i^Y$表示意图标签，$\hat{\kappa}_j^{i,C}$表示针对位置为 i 的字符的槽位标签，$\kappa_j^{i,C}$表示针对位置为 i 的字符的槽位标签的实际值，n_Y表示意图标签的数量，n_C表示槽位标签的数量；

最后结合损失函数$\mathcal{L}_Y$、$\mathcal{L}_C$、$\mathcal{L}_G$、$\mathcal{L}_{CL-C}$和$\mathcal{L}_{CL-Y}$，通过调谐线性组合构成全局语义级意图-槽位对比学习的总体损失函数：

$$\mathcal{L}_{total}=\lambda_1\mathcal{L}_Y+\lambda_2\mathcal{L}_C+\lambda_3\mathcal{L}_{GC-Y}+\lambda_4\mathcal{L}_{GC-C}+\lambda_5\mathcal{L}_{\mathrm{G}}$$

式中，λ表示可训练的参数，为标量。

最后说明的是，以上实施例仅用以说明本发明的技术方案而非限制，尽管参照较佳实施例对本发明进行了详细说明，本领域的普通技术人员应当理解，可以对本发明的技术方案进行修改或者等同替换，而不脱离本技术方案的宗旨和范围，其均应涵盖在本发明的权利要求

范围当中。

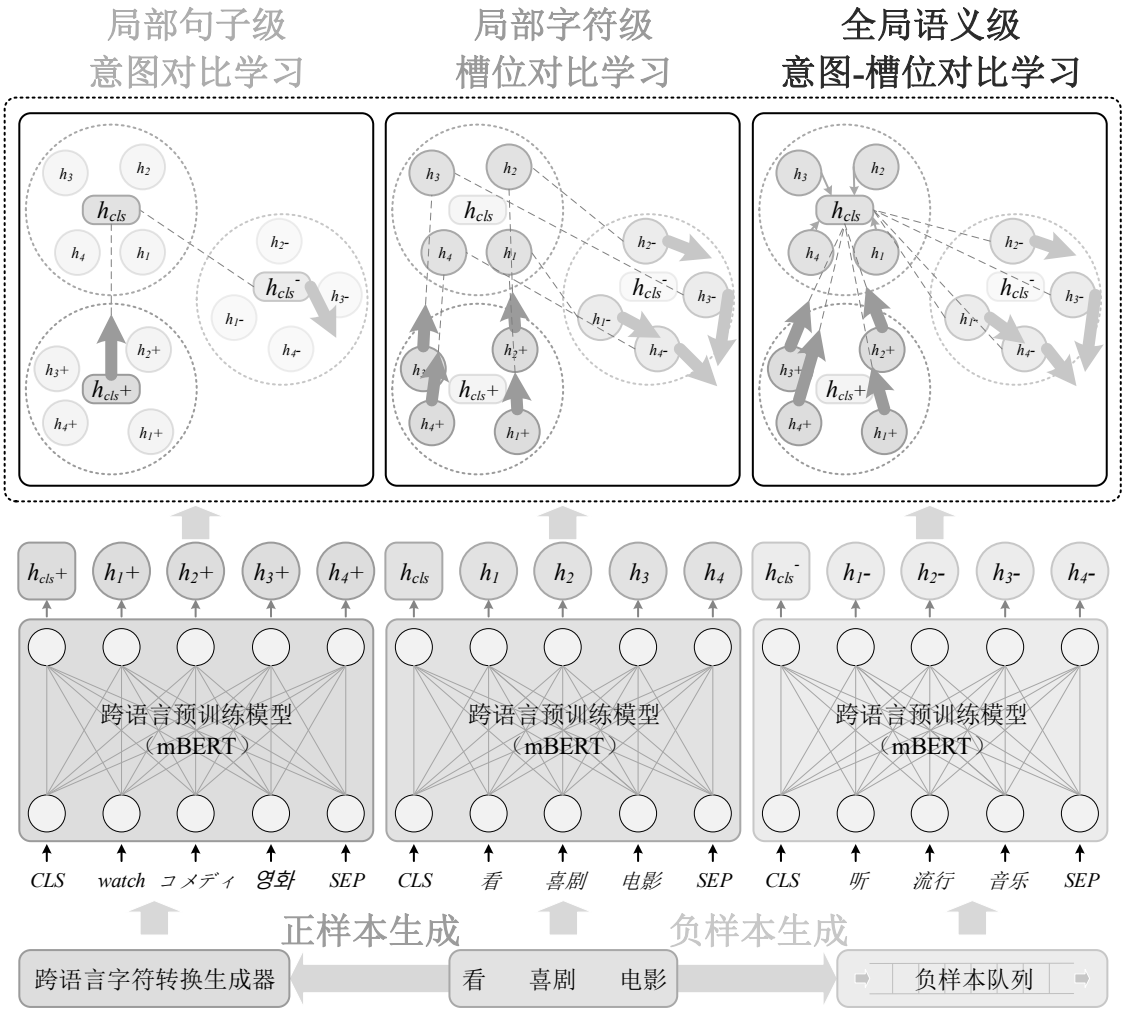

局部句子级
意图对比学习
局部字符级
槽位对比学习
全局语义级
意图-槽位对比学习
跨语言预训练模型
（mBERT）
跨语言预训练模型
（mBERT）
跨语言预训练模型
（mBERT）
CLS watch コメディ 영화 SEP
CLS 看 喜剧 电影 SEP
CLS 听 流行 音乐 SEP
正样本生成
负样本生成
跨语言字符转换生成器
看 喜剧 电影
负样本队列

图 1