

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2024 年 10 月 17 日 (17.10.2024)



(10) 国际公布号
WO 2024/215250 A1

- (51) 国际专利分类号:
G10L 15/00 (2013.01) *G06N 3/08* (2023.01)
G06F 40/00 (2020.01) *G06F 16/332* (2019.01)
G06N 20/00 (2019.01)
- (21) 国际申请号: PCT/SG2024/050199
- (22) 国际申请日: 2024 年 3 月 29 日 (29.03.2024)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202310397072.3 2023年4月11日 (11.04.2023) CN
- (71) 申请人: 阿里巴巴创新公司 (ALIBABA INNOVATION PRIVATE LIMITED) [SG/SG]; 新加坡新加坡勿拉士峇沙路 51 号来赞达一号大厦#03-06, Singapore 189554 (SG)。

(72) 发明人: 林廷恩 (LIN, Tingen); 中国北京市朝阳区望京东园四区 6 号楼金辉大厦, Beijing 100102 (CN)。余天枢 (YU, Tianshu); 中国北京市朝阳区望京东园四区 6 号楼金辉大厦, Beijing 100102 (CN)。高浩瑜 (GAO, Haoyu); 中国北京市朝阳区望京东园四区 6 号楼金辉大厦, Beijing 100102 (CN)。杨敏 (YANG, Min); 中国北京市朝阳区望京东园四区 6 号楼金辉大厦, Beijing 100102 (CN)。武玉川 (WU, Yuchuan); 中国北京市朝阳区望京东园四区 6 号楼金辉大厦, Beijing 100102 (CN)。马文涛 (MA, Wentao); 中国北京市朝阳区望京东园四区 6 号楼金辉大厦, Beijing 100102 (CN)。黄非 (HUANG, Fei); 美国加利福尼亚州森尼韦阿尔曼街 525 号 4 楼, California 94085 (US)。李永彬 (LI, Yongbin); 中国北京市朝阳区望京东园四区 6 号楼金辉大厦, Beijing 100102 (CN)。

(54) Title: MACHINE LEARNING MODEL TRAINING METHOD AND APPARATUS, DIALOGUE METHOD AND APPARATUS, DEVICE, AND MEDIUM

(54) 发明名称: 机器学习模型训练、对话方法、装置、设备及介质

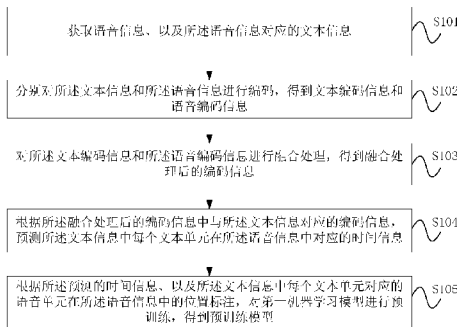


图 1

(57) Abstract: Embodiments of the present disclosure relate to a machine learning model training method and apparatus, a dialogue method and apparatus, a device, and a medium. The machine learning model training method comprises: acquiring speech information and text information corresponding to the speech information, and respectively coding the text information and the speech information to obtain text coded information and speech coded information; fusing the text coded information and the speech coded information to obtain fused coded information; according to the fused coded information, predicting time information, in the speech information, corresponding to each text unit in the text information; and according to the predicted time information and a position annotation, in the speech information, of a speech unit corresponding to each text unit in the text information, pretraining a first machine learning model to obtain a pre-trained model.

(57) 摘要: 说明书摘要本公开实施例涉及一种机器学习模型训练、对话方法、装置、设备及介质。本公开通过获取语音信息、以及语音信息对应的文本信息,并分别对文本信息和语音信息进行编码,得到文本编码信息和语音编码信息。对文本编码信息和语音编码信息进行融合处理,得到融合处理后的编码信息,根据融合处理后的编码信息,预测文本信息中每个文本单元在语音信息中对应的时间信息,根据预测的时间信息、以及文本信息中每个文本单元对应的语音单元在语音信息中的位置标注,对第一机器学习模型进行预训练,得到预训练模型。

[见续页]

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

机器学习模型训练、对话方法、装置、设备及介质

交叉援引

本公开要求于 2023 年 04 月 11 日提交中国专利局、申请号 202310397072.3、申请名称“机器学习模型训练、对话方法、装置、设备及介质”的中国专利申请的优先权，其全部内容通过引用结合在本公开中。

技术领域

本公开涉及信息技术领域，尤其涉及一种机器学习模型训练、对话方法、装置、设备及介质。

背景技术

目前，智能客服在一些业务场景下可以与用户进行对话，从而为用户提供业务咨询、业务办理等服务。具体的，智能客服可以通过部署在计算设备例如计算机或计算机集群上的机器学习模型来实现。

但是，目前的机器学习模型对用户话语理解的精准度不高。

发明内容

为了解决上述技术问题或者至少部分地解决上述技术问题，本公开提供了一种机器学习模型训练、对话方法、装置、设备及介质，以提高预训练模型对语音信息和文本信息的理解力，从而提高预训练模型对用户话语理解的精准度。

第一方面，本公开实施例提供一种机器学习模型训练方法，包括：

获取语音信息、以及所述语音信息对应的文本信息；

分别对所述文本信息和所述语音信息进行编码，得到文本编码信息和语音编码信息；

对所述文本编码信息和所述语音编码信息进行融合处理，得到融合处理后的编码信息；

根据所述融合处理后的编码信息中与所述文本信息对应的编码信息，预测所述文本信息中每个文本单元在所述语音信息中对应的时间信息；

根据所述预测的时间信息、以及所述文本信息中每个文本单元对应的语音单元在所述语音信息中的位置标注，对第一机器学习模型进行预训练。

第二方面，本公开实施例提供一种机器学习模型训练方法，所述方法包括：

获取目标任务的样本数据，所述样本数据包括待处理数据和所述待处理数据的标注结果，所述待处理数据包括样本语音、以及所述样本语音对应的样本文本；

将所述待处理数据输入到第二机器学习模型中，使得所述第二机器学习模型输出处理后的结果，所述第二机器学习模型是根据如第一方面所述的训练方法得到的预训练模型；

根据所述第二机器学习模型输出的处理后的结果、以及所述待处理数据的标注结果，对所述第二机器学习模型进行训练，得到训练后的机器学习模型。

第三方面，本公开实施例提供一种智能对话方法，所述方法包括：

获取目标用户和智能对话系统之间的历史对话，以及所述历史对话对应的文本信息；

将所述历史对话和所述文本信息输入到训练后的机器学习模型中，使得所述机器学习模型输出所述目标用户的情感、情绪、意图、关键信息中的至少一种，所述机器学习模型是根据第二方面所述的方法训练得到的；

根据所述目标用户的情感、情绪、意图、关键信息中的至少一种，与所述目标用户进行智能对话。

第四方面，本公开实施例提供一种机器学习模型训练装置，包括：

获取模块，被设置为获取语音信息、以及所述语音信息对应的文本信息；

编码模块，被设置为分别对所述文本信息和所述语音信息进行编码，得到文本编码信息和语音编码信息；

融合处理模块，被设置为对所述文本编码信息和所述语音编码信息进行融合处理，得到融合处理后的编码信息；

预测模块，被设置为根据所述融合处理后的编码信息中与所述文本信息对应的编码信息，预测所述文本信息中每个文本单元在所述语音信息中对应的时间信息；

预训练模块，被设置为根据所述预测的时间信息、以及所述文本信息中每个文本单元对应的语音单元在所述语音信息中的位置标注，对第一机器学习模型进行预训练。

第五方面，本公开实施例提供一种电子设备，包括：

存储器；

处理器；以及

计算机程序；

其中，所述计算机程序存储在所述存储器中，并被配置为由所述处理器执行以实现如第一方面、第二方面或第三方面所述的方法。

第六方面，本公开实施例提供一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行以实现第一方面、第二方面或第三方面所述的方法。

本公开实施例提供的机器学习模型训练、对话方法、装置、设备及介质，通过获取语音信息、以及所述语音信息对应的文本信息，并分别对所述文本信息和所述语音信息进行编码，得到文本编码信息和语音编码信息。进一步，对所述文本编码信息和所述语音编码信息进行融合处理，得到融合处理后的编码信息，根据所述融合处理后的编码信息，预测所述文本信息中每个文本单元在所述语音信息中对应的时间信息，然后根据所述预测的时间信息、以及所述文本信息中每个文本单元对应的语音单元在所述语音信息中的位置标注，对第一机器学习模型进行预训练，得到预训练模型。使得预训练模型可以将语音信息和文本信息在时序上进行对齐，例如，文本信息中更细粒度的文本单元与语音信息中更细粒度的语音单元之间的对齐，从而提高预训练模型对语音信息和文本信息的理解力，从而提高预训练模型对用户话语理解的精准度。

附图说明

此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本公开的实施例，并与说明书一起用于解释本公开的原理。

为了更清楚地说明本公开实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，对于本领域普通技术人员而言，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

图1为本公开实施例提供的模型预训练方法流程图；

图2为本公开实施例提供的应用场景的示意图；

图3为本公开实施例提供的机器学习模型的示意图；

图4为本公开实施例提供的用户与智能对话系统对话的示意图；

图5为本公开另一实施例提供的时序预测的示意图；

图6为本公开另一实施例提供的机器学习模型的示意图；

图7为本公开另一实施例提供的第一损失函数的构造示意图；

图8为本公开另一实施例提供的多种类型的样本数据的示意图；

图9为本公开另一实施例提供的模型预训练方法流程图；

图10为本公开另一实施例提供的机器学习模型的示意图；

图 11 为本公开另一实施例提供的模型训练方法流程图；
图 12 为本公开另一实施例提供的智能对话方法流程图；
图 13 为本公开实施例提供的模型预训练装置的结构示意图；
图 14 为本公开实施例提供的模型训练装置的结构示意图；
图 15 为本公开实施例提供的智能对话装置的结构示意图；
图 16 为本公开实施例提供的电子设备实施例的结构示意图。

具体实施方式

为了能够更清楚地理解本公开的上述目的、特征和优点，下面将对本公开的方案进行进一步描述。需要说明的是，在不冲突的情况下，本公开的实施例及实施例中的特征可以相互组合。

在下面的描述中阐述了 many 具体细节以便于充分理解本公开，但本公开还可以采用其他不同于在此描述的方式来实施；显然，说明书中的实施例只是本公开的一部分实施例，而不是全部的实施例。

需要说明的是，本公开所涉及的语音信息（包括但不限于终端设备采集的用户语音）和文本信息（包括但不限于用户语音对应的文本信息等），均为经用户授权或者经过各方充分授权的信息和数据，并且相关数据的收集、使用和处理需要遵守相关国家和地区的相关法律法规和标准，并提供有相应的操作入口，供用户选择授权或者拒绝。

另外，本公开提供的机器学习模型训练、对话方法会涉及到如下的几个术语解释，详细内容如下：

语音语义联合建模：系统同时将语音和对应的文本转写作为输入，在语音文本对齐等任务的引导下，实现语音和文本理解的共同提升。

预训练：重要的手段，让模型在大规模语料上进行无监督的学习，可以大幅度提升下游任务的结果。

时序性：指输入序列有时间顺序的特性。语音和文本都具有时序性。

导航：用户打电话给语音机器人。

外呼：语音机器人向用户拨打电话。

通常情况下，智能客服在一些业务场景下可以与用户进行对话，从而为用户提供业务咨询、业务办理等服务。具体的，智能客服可以通过部署在计算设备例如计算机或计算机集群上的机器学习模型来实现。但是，目前的机器学习模型对用户话语理解的精准度不高。针对该问题，本公开实施例提供了一种机器学习模型训练方法，下面结合具体的实施例对该方法进行介绍。

图 1 为本公开实施例提供的机器学习模型训练方法流程图。该机器学习模型训练方法具体可以是预训练方法。该方法可以由机器学习模型训练装置执行，该装置可以采用软件和/或硬件的方式实现，该装置可配置于电子设备中，例如服务器或终端中，其中，终端具体包括手机、电脑或平板电脑等。服务器具体可以是云端的服务器，该机器学习模型训练方法可以由云端来执行，在云端可以部署有若干计算节点（云服务器），每个计算节点中都具有计算、存储等处理资源。在云端，可以组织由多个计算节点来提供某种服务，当然，一个计算节点也可以提供一种或多种服务。云端提供该服务的方式可以是对外提供服务接口，用户调用该服务接口以使用相应的服务。服务接口包括软件开发工具包（Software Development Kit，简称 SDK）、应用程序接口（Application Programming Interface，简称 API）等形式。另外，本实施例所述的机器学习模型训练方法可以适用于如图 2 所示的应用场景。如图 2 所示，该应用场景包括终端 21 和服务器 22，其中，服务器 22 可以采用本实施例所述的机器学习模型训练方法对第一机器学习模型进行预训练，该第一机器学习模型经过预训练之后得到预训练模型，进一步，该预训练模型可以保留在服务器 22 中，或者服务器 22 可以将预训练

模型部署到终端 21 或其他服务器上。下面结合图 2 对该方法进行详细介绍，如图 1 所示，该方法具体步骤如下：

S101、获取语音信息、以及所述语音信息对应的文本信息。

例如，服务器 22 中可设置有智能对话系统，如图 2 所示的终端 21 可以采集用户语音，并将用户语音发送给服务器 22，该智能对话系统可以根据该用户语音进行语音回复。进一步，服务器 22 将该智能对话系统的语音回复反馈给终端 21，依此类推，使得终端 21 的用户可以与该智能对话系统进行语音对话。在对话过程中，服务器 22 可以获取语音信息，该语音信息可以是用户与智能对话系统之间的语音对话。进一步，服务器 22 还可以通过语音识别（Automatic Speech Recognition, ASR）将该语音对话转换为相应的文本信息，即该文本信息是该语音对话的文本转写。或者，服务器 22 获取的语音信息可以是用户语音，该文本信息是该用户语音的文本转写。

具体的，该语音信息、以及该语音信息对应的文本信息可作为第一机器学习模型的输入。在本实施例中，该文本信息可以包括至少一个句子。该语音信息可包括该至少一个句子中每个句子的语音，或者包括该至少一个句子中部分句子的语音。

S102、分别对所述文本信息和所述语音信息进行编码，得到文本编码信息和语音编码信息。

在本实施例中，第一机器学习模型可包括文本编码器和语音编码器，其中，文本编码器可被设置为对文本信息进行编码，得到文本编码信息。语音编码器可被设置为对语音信息进行编码，得到语音编码信息。

在一种可行的实现方式中，文本编码器可以对该文本信息中包括的每个句子分别进行编码，得到文本编码信息，该文本编码信息包括每个句子的表示向量。同理，语音编码器可以对该语音信息中每个句子的语音分别进行编码，得到语音编码信息，该语音编码信息包括每个句子的语音的表示向量。

在另一种可行的实现方式中，分别对所述文本信息和所述语音信息进行编码，得到文本编码信息和语音编码信息，包括：通过文本编码器对所述文本信息中每个文本单元分别包括的至少一个编码单元进行编码，得到文本编码信息，所述文本编码信息包括各个编码单元的表示向量；通过语音编码器对所述语音信息中的每个语音片段进行编码，得到语音编码信息，所述语音编码信息包括各个语音片段的表示向量。

例如，该文本信息中的每个句子由文本单元构成，该文本单元可以是字符、字、词、词组、子词、单词等。例如，一个句子是“他们赢了五把”，以词作为文本单元时，该句子可分割为 5 个词，分别是“他”、“们”、“赢”、“了”、“五把”。其中，每个文本单元还可以包括至少一个编码单元，例如，“他”包括一个字，则“他”包括一个编码单元，即“他”本身。“五把”包括两个字，如果将一个字作为一个编码单元，则“五把”包括两个编码单元。因此，“他们赢了五把”这句话包括 6 个编码单元，分别是“他”、“们”、“赢”、“了”、“五”、“把”。因此，通过文本编码器对“他们赢了五把”这句话编码时，可以对该 6 个编码单元分别进行编码，经过编码之后，每个编码单元对应一个表示向量。

再例如，一个句子是“*They won five games successively*”，该句子一共包括 5 个单词，每个单词作为一个文本单元，由于前 4 个单词分别较短，因此，前 4 个单词中的每个单词可分别作为一个编码单元。但是，最后一个单词即“*successively*”较长，此处可以将该单词划分为两个编码单元，分别是“*success*”和“*ively*”。因此，通过文本编码器对“*They won five games successively*”这句话编码时，可以对 6 个编码单元即“*They*”、“*won*”、“*five*”、“*games*”、“*success*”、“*ively*”分别进行编码，经过编码之后，每个编码单元对应一个表示向量。

由于该文本信息包括的句子不限于一个，每个句子可包括多个文本单元，每个文本单元还可以包括至少一个编码单元，因此，通过文本编码器对该文本信息进行编码

时,可通过文本编码器对该文本信息中的每个编码单元进行编码,得到文本编码信息,文本编码信息包括各个编码单元的表示向量。

另外,语音信息中每个句子的语音可以是长度为 10 秒的音频信号,在进行语音编码之前,还可以将每个句子的语音拆分为多个语音片段,例如,每个语音片段的长度可以是 1 秒。进一步,通过语音编码器对该语音信息中的每个语音片段分别进行编码,得到语音编码信息,该语音编码信息包括各个语音片段的表示向量。

S103、对所述文本编码信息和所述语音编码信息进行融合处理,得到融合处理后的编码信息。

在本实施例中,第一机器学习模型还可以包括模态融合层。在得到如上所述的文本编码信息和语音编码信息之后,可以将该文本编码信息和语音编码信息进行拼接,并将拼接后的结果作为该模态融合层的输入,使得该模态融合层对该文本编码信息和语音编码信息进行融合处理,得到融合处理后的编码信息,该融合处理后的编码信息也可记为联合表示向量。具体的,模态融合层的输入和输出一一对应,例如,模态融合层的输入包括 m 个编码单元的表示向量和 n 个语音片段的表示向量,即模态融合层的输入包括 $m+n$ 个表示向量,则模态融合层的输出也包括 $m+n$ 个表示向量,并且输入中的 $m+n$ 个表示向量和输出中的 $m+n$ 个表示向量一一对应。对于模态融合层输出的每个表示向量而言,该表示向量同时包含文本成分和语音成分。因此,模态融合层输出的每个表示向量可记为融合表示向量。

S104、根据所述融合处理后的编码信息中与所述文本信息对应的编码信息,预测所述文本信息中每个文本单元在所述语音信息中对应的时间信息。

例如,文本信息包括 m 个编码单元,语音信息包括 n 个语音片段,模态融合层的输入包括 m 个编码单元的表示向量和 n 个语音片段的表示向量。由于模态融合层的输入和输出是一一对应的,因此,模态融合层输出的前 m 个表示向量是与该文本信息对应的编码信息,模态融合层输出的后 n 个表示向量是与该语音信息对应的编码信息。由于模态融合层输出的每个表示向量同时包含文本成分和语音成分,因此,第一机器学习模型根据模态融合层输出的前 m 个表示向量,可预测出该文本信息中每个文本单元在该语音信息中对应的时间信息。例如,每个文本单元在该语音信息中对应的时间信息包括该文本单元在该语音信息中对应的开始时间和结束时间。可以理解的是,文本单元例如“他”、“们”、“赢”、“了”、“五把”分别是文本类型的信息,每个文本单元分别对应一个语音词,该语音词是语音类型的信息,一个文本单元对应的语音词可以是用户或智能对话系统对该文本单元的发音。因此,该文本单元在该语音信息中对应的开始时间具体是该文本单元对应的语音词在该语音信息中对应的开始时间(例如发音的开始时刻),该文本单元在该语音信息中对应的结束时间具体是该文本单元对应的语音词在该语音信息中对应的结束时间(例如发音的结束时刻)。

S105、根据所述预测的时间信息、以及所述文本信息中每个文本单元对应的语音单元在所述语音信息中的位置标注,对第一机器学习模型进行预训练,得到预训练模型。

具体的,在本实施例中,可以预先标注该文本信息中的每个文本单元对应的语音单元在该语音信息中的位置。进一步,根据第一机器学习模型预测出的文本信息中每个文本单元在所述语音信息中对应的时间信息、以及预先标注该文本信息中的每个文本单元对应的语音单元在该语音信息中的位置,构建第一损失函数,然后根据该第一损失函数对该第一机器学习模型进行预训练。

例如,本实施例中的文本信息是“他们赢了五把”,语音信息是“他们赢了五把”的语音。在本实施例中,可以预先标注该文本信息中的每个文本单元对应的语音单元在该语音信息中的位置。例如,该文本信息包括“他”、“们”、“赢”、“了”、“五把”这 5 个文本单元。该语音信息包括“他”、“们”、“赢”、“了”、“五

把”分别对应的语音词，该语音词记为语音单元，例如，“他”对应的语音词是“他”的发音。本实施例可以预先标注“他”对应的语音词在该语音信息中的开始位置（例如开始时间或开始时刻）和结束位置（例如结束时间或结束时刻）。同理，可以预先标注“们”、“赢”、“了”、“五把”分别对应的语音词在该语音信息中的开始位置和结束位置。进一步，根据第一机器学习模型预测出的“他”、“们”、“赢”、“了”、“五把”分别在该语音信息中的开始时间和结束时间、以及预先标注的“他”、“们”、“赢”、“了”、“五把”分别在该语音信息中的开始位置和结束位置，构建第一损失函数。

具体的，根据第一机器学习模型预测出的“他”、“们”、“赢”、“了”、“五把”分别在该语音信息中的开始时间、以及预先标注的“他”、“们”、“赢”、“了”、“五把”分别对应的语音词在该当前语音中的开始位置，计算第一差值。根据第一机器学习模型预测出的“他”、“们”、“赢”、“了”、“五把”分别在该语音信息中的结束时间、以及预先标注的“他”、“们”、“赢”、“了”、“五把”分别对应的语音词在该当前语音中的结束位置，计算第二差值。进一步，根据第一差值和第二差值构建第一损失函数。并根据该第一损失函数对第一机器学习模型进行预训练，得到预训练模型。

本公开实施例通过获取语音信息、以及所述语音信息对应的文本信息，并分别对所述文本信息和所述语音信息进行编码，得到文本编码信息和语音编码信息。进一步，对所述文本编码信息和所述语音编码信息进行融合处理，得到融合处理后的编码信息，根据所述融合处理后的编码信息，预测所述文本信息中每个文本单元在所述语音信息中对应的时间信息，然后根据所述预测的时间信息、以及所述文本信息中每个文本单元对应的语音单元在所述语音信息中的位置标注，对第一机器学习模型进行预训练，得到预训练模型。使得预训练模型可以将语音信息和文本信息在时序上进行对齐，例如，文本信息中更细粒度的文本单元与语音信息中更细粒度的语音单元之间的对齐，从而提高预训练模型对语音信息和文本信息的理解力，从而提高预训练模型对用户话语理解的精准度。

如图3所示为第一机器学习模型的结构示意图。具体的，该第一机器学习模型包括文本编码器、语音编码器、模态融合层。下面结合图3对预训练方法进行详细介绍。

具体的，第一机器学习模型的输入包括文本信息和语音信息，即第一机器学习模型的输入是多模态信息。可选的，所述语音信息包括历史语音和当前语音，所述文本信息包括历史文本和当前文本。也就是说，对于每一种模态的信息而言，均包括历史信息 and 当前信息。例如，当该语音信息是用户与智能对话系统之间的语音对话时，随着时间的增加，对话内容在不断的增加，因此，在对话过程中，有些语音可作为历史语音，有些语音可作为当前语音。同理，由于文本信息是该语音对话的文本转写，因此，文本信息中有些句子可作为历史文本，有些句子可作为当前文本。

如图4所示为用户与智能对话系统之间的对话示意图。假设图4所示的对话内容是文本转写。用户说的每句话的文本转写可记为一轮文本，智能对话系统说的每句话的文本转写可记为一轮文本，因此，如图4所示的对话内容是8轮文本，即“AAAAAA”、“BBBBBB”、“CCCC”、“DDDD”、“EEEEEE”、“FFFFFF”、“GGGGG”、“HHHHH”分别是一轮文本。另外，本实施例可以将该8轮文本中的前7轮文本记为历史文本，最后1轮文本即文本“HHHHH”记为当前文本。同理，由于8轮文本中的每轮文本对应一个语音，且一个语音记为一轮语音，因此，8轮文本对应8轮语音，该8轮语音中的前7轮语音记为历史语音，最后1轮语音即“HHHHH”对应的语音记为当前语音。此外，每一轮语音的最大长度可以是10秒。

由于文本信息和语音信息是两种不同模态的信息，因此，同一轮数的文本和同一轮数的语音所占的存储空间可能是不同的，例如，同一轮数的语音所占的存储空间可

能更大，因此，可以将同一轮数的文本和同一轮数的语音作为第一机器学习模型的输入，或者将第一轮数的文本和第二轮数的语音作为第一机器学习模型的输入，此处第一轮数大于第二轮数。

例如，在本实施例中，可以选择8轮文本和2轮语音作为如图3所示的第一机器学习模型的输入。具体的，该8轮文本中包括7轮历史文本和1轮当前文本。该2轮语音中包括1轮历史语音和1轮当前语音。另外，此处的1轮当前文本和1轮当前语音是对齐的，例如，此处的1轮当前文本是“HHHHH”，1轮当前语音是“HHHHH”对应的语音。因此，如图3所示的历史文本可以是此处的7轮历史文本，如图3所示的当前文本可以是此处的1轮当前文本。如图3所示的历史语音可以是此处的1轮历史语音，如图3所示的当前语音可以是此处的1轮当前语音。即7轮历史文本和1轮当前文本构成文本信息，1轮历史语音和1轮当前语音构成语音信息。进一步，通过文本编码器对该文本信息进行编码，得到文本编码信息；通过语音编码器对该语音信息进行编码，得到语音编码信息。

如图3所示，为了将7轮历史文本和1轮当前文本进行区分，可以在7轮历史文本的起始位置增加预设符号<s>，在7轮历史文本和1轮当前文本之间的位置增加预设符号</s>。或者，在每相邻的两轮文本之间的位置增加预设符号</s>，即</s>被设置为分割每一轮文本。同理，可以采用类似于<s>和</s>的预设符号，将历史语音和当前语音分割开，或者将每一轮语音分割开。

进一步，通过文本编码器对<s>、如上所述的7轮历史文本、</s>和如上所述的1轮当前文本进行编码。通过语音编码器对如上所述的1轮历史语音和1轮当前语音进行编码。

例如，对于如上所述的7轮历史文本和1轮当前文本而言，每一轮文本可以是一个句子。每个句子由文本单元构成，每个文本单元还可以包括至少一个编码单元。因此，可以将如图3所示的历史文本和当前文本分别分割为若干个编码单元，例如框31表示任意一个编码单元。同时，预设符号<s>可作为一个编码单元，预设符号</s>可作为一个编码单元。如图3所示，通过文本编码器对各个编码单元分别进行编码之后，得到文本编码信息，该文本编码信息包括各个编码单元的表示向量。例如，该文本编码信息包括如图3所示的框32内的各个表示向量，框32内的每个小框（例如框33）表示一个编码单元的表示向量。

另外，在本实施例中，如图3所示的历史语音和当前语音分别是长度为10秒的音频信号，在进行编码之前，可以将该历史语音拆分为多个语音片段，将该当前语音也拆分为多个语音片段，例如，每个语音片段的长度可以是1秒。如图3所示，通过语音编码器对该历史语音中的每个语音片段进行编码，以及通过语音编码器对该当前语音中的每个语音片段进行编码，得到语音编码信息，该语音编码信息包括各个语音片段的表示向量。例如，该语音编码信息包括如图3所示的框34内的各个表示向量，框34内的每个小框（例如框35）表示一个语音片段的表示向量。

如图3所示，将框32表示的文本编码信息和框34表示的语音编码信息进行拼接，并将拼接后的结果作为模态融合层的输入，使得模态融合层对文本编码信息和语音编码信息进行融合处理，得到融合处理后的编码信息，该融合处理后的编码信息也可记为联合表示向量。例如，该融合处理后的编码信息包括如图3所示的框36内的各个表示向量，框36内的每个小框（例如框37）表示模态融合层输出的一个表示向量。例如框33表示预设字符<s>的表示向量，框37表示模态融合层输出的第一个表示向量，该第一个表示向量是文本编码信息和语音编码信息的融合结果，即该第一个表示向量是预设字符<s>的表示向量、历史文本中每个编码单元的表示向量、预设字符</s>的表示向量、当前文本中每个编码单元的表示向量、历史语音中每个语音片段的表示向量、当前语音中每个语音片段的表示向量的融合结果，因此，第一个表示向量中不

仅包含文本成分，还包括语音成分。同理，模态融合层输出的其他表示向量均同时包含文本成分和语音成分。

进一步，从图 3 所示的框 36 内的各个表示向量中截取部分表示向量，例如截取模态融合层输出的与当前文本对应的表示向量，即框 38 内的各个表示向量。进一步，根据框 38 内的各个表示向量，预测当前文本中每个文本单元在当前语音中对应的时间信息。例如，当前语音是一个 10 秒长的语音。对该当前语音的长度进行归一化处理后，可将当前语音的开始位置记为 0，结束位置记为 1。如图 3 所示，通过全连接层 1 预测该当前文本中每个文本单元在当前语音中对应的开始时间，通过全连接层 2 预测该当前文本中每个文本单元在当前语音中对应的结束时间。例如图 5 所示，假设当前文本是“他们赢了五把”，根据模态融合层输出的与当前文本对应的表示向量，可预测出“他”在当前语音中对应的开始时间（例如 0.05）和结束时间（例如 0.16），“们”在当前语音中对应的开始时间（例如 0.18）和结束时间（例如 0.32），依次类推，“五把”在当前语音中对应的开始时间（例如 0.067）和结束时间（例如 0.97）。

例如，当前文本和当前语音分别是样本，因此，可以预先标注当前文本中的每个文本单元对应的语音单元在该语音信息中的位置。例如，该当前语音中包括“他”、“们”、“赢”、“了”、“五把”分别对应的语音词，该语音词记为语音单元，例如，“他”对应的语音词是“他”的发音。本实施例可以预先标注“他”对应的语音词在该当前语音中的开始位置（例如开始时间或开始时刻）和结束位置（例如结束时间或结束时刻）。同理，可以预先标注“们”、“赢”、“了”、“五把”分别对应的语音词在该当前语音中的开始位置和结束位置。进一步，根据全连接层 1 预测的“他”、“们”、“赢”、“了”、“五把”分别对应的语音词在该当前语音中的开始位置，构建第一差值。根据全连接层 2 预测的“他”、“们”、“赢”、“了”、“五把”分别对应的语音词在该当前语音中的结束位置，构建第二差值。进一步，根据第一差值和第二差值构建第一损失函数，通过该第一损失函数对如图 3 所示的第一机器学习模型进行预训练，得到预训练模型。

在上述实施例中，所述融合处理后的编码信息包括所述各个编码单元分别对应的融合表示向量、以及所述各个语音片段分别对应的融合表示向量；根据所述融合处理后的编码信息，预测所述文本信息中每个文本单元在所述语音信息中对应的时间信息，包括：若所述文本单元包括一个编码单元，则根据所述编码单元对应的融合表示向量，预测所述文本单元在所述语音信息中对应的开始时间和结束时间；若所述文本单元包括多个编码单元，则根据所述多个编码单元中第一个编码单元对应的融合表示向量，预测所述文本单元在所述语音信息中对应的开始时间，以及根据所述多个编码单元中最后一个编码单元对应的融合表示向量，预测所述文本单元在所述语音信息中对应的结束时间。

如图 6 所示，模态融合层输出的各个融合表示向量中包括预设字符<s>对应的融合表示向量、历史文本中各个编码单元分别对应的融合表示向量、预设字符</s>对应的融合表示向量、当前文本中各个编码单元分别对应的融合表示向量、以及各个语音片段分别对应的融合表示向量。具体的，如果历史文本有 8 轮，历史语音有 8 轮，并且 8 轮历史文本和 8 轮历史语音是一一对齐的，那么根据历史文本中各个编码单元分别对应的融合表示向量，可以预测历史文本中每个文本单元在历史语音中对应的时间信息。例如，根据第 1 轮历史文本中各个编码单元分别对应的融合表示向量，预测该第 1 轮历史文本中每个文本单元在第 1 轮历史语音中对应的开始时间和结束时间，依此类推，预测该第 8 轮历史文本中每个文本单元在第 8 轮历史语音中对应的开始时间和结束时间。另外，由于当前文本和当前语音是对齐的，因此，可以根据当前文本中

各个编码单元分别对应的融合表示向量,预测当前文本中每个文本单元在当前语音中对应的开始时间和结束时间。

例如,当前文本是“他们赢了五把”,“他”、“们”、“赢”、“了”、“五”、“把”分别是编码单元。模态融合层输出的各个融合表示向量中包括“他”、“们”、“赢”、“了”、“五”、“把”分别对应的融合表示向量。例如图6所示的A表示“他”对应的融合表示向量,依次类推,F表示“把”对应的融合表示向量。由于“他”、“们”、“赢”、“了”分别是文本单元,并且该文本单元包括一个编码单元,例如,“他”是一个文本单元,并且该文本单元包括一个编码单元,该编码单元是“他”,即“他”既是一个文本单元,也是该文本单元中包括的一个编码单元,因此,可以根据“他”对应的融合表示向量A,预测“他”在当前语音中对应的开始时间和结束时间。例如图6所示的全连接层1包括参数 W_{start} ,全连接层2包括参数 W_{end} ,由全连接层1预测的“他”在当前语音中对应的开始时间记为 $W_{start}A$,由全连接层2预测的“他”在当前语音中对应的结束时间记为 $W_{end}A$ 。同理,全连接层1还可以预测出“们”、“赢”、“了”分别在当前语音中对应的开始时间,全连接层2还可以预测出“们”、“赢”、“了”分别在当前语音中对应的结束时间。

另外,由于“五把”是文本单元,但是该文本单元包括多个编码单元,例如“五”是该文本单元中的第一个编码单元,“把”是该文本单元中的最后一个编码单元。因此,根据第一个编码单元对应的融合表示向量,可预测“五把”在当前语音中对应的开始时间,以及根据最后一个编码单元对应的融合表示向量,可预测“五把”在当前语音中对应的结束时间。例如,“五”对应的融合表示向量记为E,“把”对应的融合表示向量记为F,则由全连接层1预测的“五把”在当前语音中对应的开始时间记为 $W_{start}E$,由全连接层2预测的“五把”在当前语音中对应的结束时间记为 $W_{end}F$ 。例如图5所示为全连接层1预测出的各个文本单元的开始时间、以及全连接层2预测出的各个文本单元的结束时间。

可选的,根据所述预测的时间信息、以及所述文本信息中每个文本单元对应的语音单元在所述语音信息中的位置标注,对第一机器学习模型进行预训练,包括:根据所述文本单元在所述语音信息中对应的开始时间和结束时间,以及预先标注的所述文本单元对应的语音单元在所述语音信息中的开始位置和结束位置,构建第一损失函数;根据所述第一损失函数,对第一机器学习模型进行预训练。可选的,所述第一损失函数包括第一差值和第二差值,所述第一差值是根据所述开始时间和所述开始位置得到的差值,所述第二差值是根据所述结束时间和所述结束位置得到的差值。

例如图7所示,当前语音的时长是10秒,即当前语音从第0秒开始、在第10秒结束。本实施例可以预先标注当前文本例如“他们赢了五把”中各个文本单元对应的语音单元在该当前语音中的开始位置和结束位置。例如,“他”对应的语音单元在该当前语音中的开始位置是第0.6秒、结束位置是1.7秒,依此类推,“五把”对应的语音单元在该当前语音中的开始位置是第6.9秒、结束位置是第9.9秒。

由于如上所述的全连接层1预测的开始时间和全连接层2预测的结束时间均是0-1之间的数值,因此,为了与该0-1之间的数值进行数量级上的匹配,本实施例可以对当前语音的时长进行归一化处理,例如,将当前语音的时长归一化为1秒,即将该当前语音的开始位置归一化为第0秒,将该当前语音的结束位置归一化为第1秒。从而使得“他们赢了五把”中各个文本单元对应的语音单元在该当前语音中的开始位置和结束位置均被归一化,例如,经过归一化处理后,“他”对应的语音单元在该当前语音中的开始位置是第0.06秒、结束位置是第0.17秒,依此类推,“五把”对应的语音单元在该当前语音中的开始位置是第0.69秒、结束位置是第0.99秒。也就是说,在归一化处理的过程中,统一除以当前语音的时长(例如10秒)。

如图7所示,针对当前文本中的同一个文本单元,例如“他”,全连接层1预测

的“他”在当前语音中对应的开始时间是 0.05，而预先标注的“他”对应的语音单元在该当前语音中的开始位置是第 0.06 秒，即该开始时间和该开始位置之间存在一定的差值，该差值记为第一差值。同时，全连接层 2 预测的“他”在当前语音中对应的结束时间是 0.16，而预先标注的“他”对应的语音单元在该当前语音中的结束位置是第 0.17 秒，即该结束时间和该结束位置之间存在一定的差值，该差值记为第二差值。另外，对于当前文本中的其他文本单元而言，也存在类似的第一差值和第二差值。此外，本实施例是以预测当前文本中各个文本单元的开始时间和结束时间为例进行的示意性说明，在其他实施例中，还可以预测历史文本中每个文本单元在历史语音中对应的的时间信息。例如，根据第 1 轮历史文本中各个编码单元分别对应的融合表示向量，预测该第 1 轮历史文本中每个文本单元在第 1 轮历史语音中对应的开始时间和结束时间，依此类推，预测该第 8 轮历史文本中每个文本单元在第 8 轮历史语音中对应的开始时间和结束时间。因此，对于每轮历史文本中的各个文本单元而言，也存在类似的第一差值和第二差值。

因此，本实施例可以基于多轮文本（包括历史文本和当前文本）中各个文本单元分别对应的第一差值和第二差值构建第一损失函数。进一步，根据该第一损失函数对如图 3 所示的第一机器学习模型进行预训练。例如，该第一损失函数 $\mathcal{L}_{TPP}$ 可表示为如下公式：

$$\mathcal{L}_{TPP} = \sum_{i=1}^m \sum_{j=1}^n \frac{1}{2} \left(\left(W_{start} h_{s_{ij}} - \frac{s_{ij}}{L_a} \right)^2 + \left(W_{end} h_{e_{ij}} - \frac{e_{ij}}{L_a} \right)^2 \right)$$

其中，该第一损失函数使用均方误差（Mean Squared Error, MSE），即预测结果（例如预测的开始时间或结束时间）与真实位置（例如预先标注的开始位置或结束位置）的差的平方。 i 表示第 i 轮文本， m 表示文本的轮数或个数。 j 表示第 i 轮文本中的第 j 个文本单元，例如第 j 个词。 n 表示第 i 轮文本中包括的词个数。 W_{start} 表示全连接层 1 包括的参数， W_{end} 表示全连接层 2 包括的参数， $h_{s_{ij}}$ 表示第 i 轮文本中第 j 个文本单元中第一个编码单元对应的融合表示向量， $h_{e_{ij}}$ 表示第 i 轮文本中第 j 个文本单元中最后一个编码单元对应的融合表示向量。如果第 j 个文本单元包括一个编码单元，则此处的第一个编码单元和最后一个编码单元是同一个编码单元，即 $h_{s_{ij}}$ 和 $h_{e_{ij}}$ 相同。如果第 j 个文本单元包括多个编码单元，则此处的第一个编码单元和最后一个编码单元不同，即 $h_{s_{ij}}$ 和 $h_{e_{ij}}$ 不同。可以理解的是，在预训练的过程中， W_{start} 和 W_{end} 分别是待训练的参数。 $W_{start} h_{s_{ij}}$ 表示全连接层 1 预测的第 i 轮文本中第 j 个文本单元在第 i 轮语音中的开始时间， $W_{end} h_{e_{ij}}$ 表示全连接层 2 预测的第 i 轮文本中第 j 个文本单元在第 i 轮语音中的结束时间。 s_{ij} 表示归一化之前，预先标注的第 i 轮文本中第 j 个文本单元在第 i 轮语音中的开始位置， e_{ij} 表示归一化之前，预先标注的第 i 轮文本中第 j 个文本单元在第 i 轮语音中的结束位置。 L_a 表示第 i 轮语音的时长，或者表示多轮语音中的最大时长， L_a 被设置为进行如上所述

的归一化处理。

另外,根据该第一损失函数对如图3或图6所示的第一机器学习模型进行预训练时,可以对如图3或图6所示的各个待训练结构,例如文本编码器、语音编码器、模态融合层、全连接层1、全连接层2中的参数进行迭代更新,直到迭代次数达到预设次数,或者参数收敛。

例如,根据该第一机器学习模型输出的预测结果和预先标注的真实位置之间的相似度或差值,调整如图3或图6所示的各个待训练结构中的参数,使得该第一机器学习模型输出的预测结果和预先标注的真实位置在后续的迭代训练过程中逐渐的相似。示例性地,可根据该第一机器学习模型输出的预测结果和预先标注的真实位置之间的梯度数据,引导参数调整。其中,梯度的本意是一个向量(矢量),表示某一函数在该点处的方向导数沿着该方向取得最大值,即函数在该点处沿着该方向(此梯度的方向)变化最快,变化率最大。也正是基于此原理,可以引导参数的调整方向,使得该第一机器学习模型输出的预测结果向预先标注的真实位置靠近。

如上所述,对第一机器学习模型进行预训练时采用的样本数据包括文本信息和语音信息。例如,该文本信息包括7轮历史文本和1轮当前文本,该语音信息包括1轮历史语音和1轮当前语音,且7轮历史文本和1轮当前文本是同一对话中连续的8个句子,同理1轮历史语音和1轮当前语音是该同一对话中连续的2个语音。即7轮历史文本和1轮当前文本之间存在连贯的语义,1轮历史语音和1轮当前语音之间存在连贯的语义。在其他实施例中,根据该样本数据可以衍生出另外三种样本数据,例如,将该样本数据中的1轮当前文本替换为其他对话中的一个句子后得到第一种衍生的样本数据,从而导致7轮历史文本和1轮当前文本之间的语义不再连贯,但1轮历史语音和1轮当前语音之间保留连贯的语义。再例如,将该样本数据中的1轮当前语音替换为其他对话中的一个语音后得到第二种衍生的样本数据,从而导致1轮历史语音和1轮当前语音之间的语义不再连贯,但7轮历史文本和1轮当前文本之间保留连贯的语义。又例如,将该样本数据中的1轮当前文本替换为其他对话中的一个句子、且将该样本数据中的1轮当前语音替换为其他对话中的一个语音后得到第三种衍生的样本数据,从而导致7轮历史文本和1轮当前文本之间的语义不再连贯,1轮历史语音和1轮当前语音之间的语义也不再连贯。如果将原始的样本数据记为如图8所示的标签(Label)为0的样本数据,那么根据该原始的样本数据衍生出的另外三种样本数据,依次是图8所示的Label为1、Label为2、Label为3的样本数据。其中, $Query_{t-1}^T$ 表示历史文本, $Query_t^T$ 表示当前文本, $Query_{t-1}^S$ 表示历史语音, $Query_t^S$ 表示当前语音, $Query_{rand}^T$ 表示替换成其他句子之后的当前文本, $Query_{rand}^S$ 表示替换成其他语音之后的当前语音。

在对第一机器学习模型进行预训练的过程中,该预训练方法还包括如图9所示的如下几个步骤:

S901、根据所述融合处理后的编码信息,判断所述历史语音和所述当前语音是否匹配,得到第一判断结果,以及判断所述历史文本和所述当前文本是否匹配,得到第二判断结果。

本实施例可以将Label为0、Label为1、Label为2、或Label为3的样本数据作为第一机器学习模型的输入。也就是说,如图10所示的当前文本可能是 $Query_t^T$ 或 $Query_{rand}^T$ 。同理,如图10所示的当前语音可能是 $Query_t^S$ 或 $Query_{rand}^S$ 。经过该第一机器学习模型中文本编码器、语音编码器、模态融合层的处理之后,模态融合层输出多个融合表示向量。本实施例可以将模态融合层输出的第一个融合表示向量输入到全连接层3,该全连接层3被设置为判断输入中的历史语音和当前语音是否匹配,即历史语音和当前语音之间的语义是否连贯,得到第一判断结果,以及判断输入中的历史文本和当前文本是否匹配,即历史文本和当前文本之间的语义是否连贯,得到第二判断结

果。

可以理解的是,如果当前文本是 $Query_i^T$,那么历史文本和当前文本匹配,即历史文本和当前文本之间的语音是连贯的,历史文本和当前文本在语义内容上具有明显的上下文关联性,此时的历史文本和当前文本是同一对话中的上下文。如果当前文本是 $Query_{rand}^T$,那么历史文本和当前文本不匹配,即历史文本和当前文本不是同一对话中的上下文。同理,如果当前语音是 $Query_i^S$,那么历史语音和当前语音匹配,即历史语音和当前语音之间的语音是连贯的,历史语音和当前语音在语义内容上具有明显的上下文关联性,此时的历史语音和当前语音是同一对话中的上下文。如果当前语音是 $Query_{rand}^S$,那么历史语音和当前语音不匹配,即历史语音和当前语音不是同一对话中的上下文。

因此,如上所述的第一判断结果有两种结果,一种结果是历史语音和当前语音匹配,另一种结果是历史语音和当前语音不匹配。同理,第二判断结果也有两种结果,一种结果是历史文本和当前文本匹配,另一种结果是历史文本和当前文本不匹配。

S902、根据所述第一判断结果和所述第二判断结果,预测输入样本的类型,所述输入样本包括所述语音信息和所述文本信息。

例如,全连接层3可以根据第一判断结果和第二判断结果预测输入样本的类型,例如,若第一判断结果表示历史语音和当前语音匹配,第二判断结果表示历史文本和当前文本匹配,则全连接层3预测输入样本是Label为0的样本。若第一判断结果表示历史语音和当前语音匹配,第二判断结果表示历史文本和当前文本不匹配,则全连接层3预测输入样本是Label为1的样本。若第一判断结果表示历史语音和当前语音不匹配,第二判断结果表示历史文本和当前文本匹配,则全连接层3预测输入样本是Label为2的样本。若第一判断结果表示历史语音和当前语音不匹配,第二判断结果表示历史文本和当前文本不匹配,则全连接层3预测输入样本是Label为3的样本。即全连接层3可以根据第一判断结果和第二判断结果预测输入样本是Label为0、Label为1、Label为2、或Label为3中的哪一种样本。

在一种可能的实现方式中,全连接层3可以直接输出其预测得到的输入样本的类型,例如,Label=0、Label=1、Label=2、或Label=3。

在另一种可能的实现方式中,全连接层3可以是一个四分类任务,即全连接层3的输出可以是4个概率值,该4个概率值依次表示输入样本的类型是Label为0、Label为1、Label为2、Label为3的概率。进一步,将该4个概率值中最大概率值对应的类型作为全连接层3预测出的输入样本的类型。

S903、根据所述预测的输入样本的类型、以及预先标注的所述输入样本的类型,构建第二损失函数,并根据所述第二损失函数对所述第一机器学习模型进行预训练。本实施例还可以预先标注输入样本中的历史语音和当前语音实际是否构成上下文信息、以及预先标注输入样本中的历史文本和当前文本实际是否构成上下文信息。即本实施例可以预先标注输入样本实际的类型,即预先标注输入样本是Label为0、Label为1、Label为2、或Label为3中的哪一种样本。

进一步,本实施例可以根据全连接层3预测的输入样本的类型、以及预先标注的输入样本实际的类型,构建第二损失函数,该第二损失函数可以是分类的交叉熵损失函数。进一步,根据该第二损失函数对如图10所示的第一机器学习模型进行预训练。

本实施例通过所述融合处理后的编码信息,判断所述历史语音和所述当前语音是否匹配,得到第一判断结果,以及判断所述历史文本和所述当前文本是否匹配,得到第二判断结果。使得第一机器学习模型可以根据第一判断结果和第二判断结果预测输入样本的类型。另外,根据所述历史语音和所述当前语音实际是否构成上下文信息、以及所述历史文本和所述当前文本实际是否构成上下文信息,预先标注输入样本的类型。从而根据预先标注的输入样本的类型、以及第一机器学习模型预测的输入样本的

类型,构建第二损失函数,并根据第二损失函数对所述第一机器学习模型进行预训练,得到预训练模型。使得预训练模型可以识别上下文信息,有利于提高预训练模型对用户意图的识别、情感分析等的精准度。

可以理解的是,本实施例可以根据第一损失函数或第二损失函数单独对第一机器学习模型进行预训练,或者可以根据第一损失函数和第二损失函数联合对第一机器学习模型进行预训练。此外,对第一机器学习模型进行预训练的任务除了如上所述的时序预测任务(例如预测文本单元在语音中的开始时间和结束时间)、跨模态对话回复选择任务(例如预测输入样本的类型)之外,还可以通过其他任务来进行预训练。例如,还可以通过掩码预测任务来对第一机器学习模型进行预训练。例如,本实施例可以对历史文本或当前文本中的部分词进行掩码,和/或,对历史语音或当前语音中的部分语音片段进行掩码。然后将掩码之后的样本输入到第一机器学习模型中,进一步,根据模态融合层输出的融合表示向量来预测被掩盖的部分词和/或部分语音。从而根据预测出的被掩盖的部分词和/或部分语音、以及实际被掩盖的部分词和/或部分语音构建第三损失函数,并根据该第三损失函数对第一机器学习模型进行预训练。从而利用不同的模态信息,例如,文本和语音,建模文本和语音两种模态的互动。

可以理解的是,对第一机器学习模型进行预训练之后可以得到预训练模型,该预训练模型可作为模型基础,进一步采用下游任务的样本数据,对该预训练模型进行训练或微调,得到训练后的机器学习模型,从而使得训练后的机器学习模型可以执行下游任务。

图 11 为本公开另一实施例提供的机器学习模型训练方法流程图。该机器学习模型训练方法可以是对该预训练模型进行训练或微调的方法。在本实施例中,该方法具体步骤如下:

S1101、获取目标任务的样本数据,所述样本数据包括待处理数据和所述待处理数据的标注结果,所述待处理数据包括样本语音、以及所述样本语音对应的样本文本。

例如,该预训练模型可以保留在服务器 22 中,进一步,服务器 22 可以对该预训练模型进行训练或微调。具体的,服务器 22 可以获取目标任务的样本数据,该目标任务可以是如上所述的下游任务,该样本数据包括待处理数据和该待处理数据的标注结果,该待处理数据包括样本语音、以及该样本语音对应的样本文本。该待处理数据的标注结果是根据该样本语音和/或样本文本,预先标注的用户情感类别、用户情绪类别、用户意图、或该样本语音或样本文本中的关键信息。

可选的,所述目标任务包括如下至少一种:情感分析任务、情绪识别任务、意图分析任务、关键信息收集任务。

例如,当目标任务是情感分析任务时,服务器 22 获取的样本数据包括样本语音、该样本语音对应的样本文本、以及预先标注的该样本语音或样本文本对应的用户情感类别。具体的,样本文本中包括历史文本和当前文本,样本语音中包括历史语音和当前语音。

S1102、将所述待处理数据输入到第二机器学习模型中,使得所述第二机器学习模型输出处理后的结果,所述第二机器学习模型是根据如上所述的预训练方法得到的预训练模型。

例如,将该历史文本、当前文本、历史语音和当前语音输入到第二机器学习模型即预训练模型中,使得该预训练模型中的文本编码器、语音编码器、模态融合层对该预训练模型的输入进行处理,并根据模态融合层输出的融合表示向量(例如第一个融合表示向量)预测该样本语音或样本文本对应的用户情感类别。该用户情感类别可作为该预训练模型输出的处理结果。

S1103、根据所述第二机器学习模型输出的处理后的结果、以及所述待处理数据的标注结果,对所述第二机器学习模型进行训练,得到训练后的机器学习模型。

例如，根据该预训练模型预测的该样本语音或样本文本对应的用户情感类别、以及预先标注的该样本语音或样本文本对应的用户情感类别，对该预训练模型进行训练或微调，从而得到训练后的机器学习模型。

同理，可以采用情绪识别任务、意图分析任务、或关键信息收集任务的样本数据对该预训练模型进行训练或微调。可以理解的是，本实施例可以通过情感分析任务、情绪识别任务、意图分析任务或关键信息收集任务单独对该预训练模型进行训练或微调。或者，通过情感分析任务、情绪识别任务、意图分析任务和关键信息收集任务联合对该预训练模型进行训练或微调。此外，不限于如上所述的几种任务对该预训练模型进行训练或微调，例如，还可以采用其他任务，对该预训练模型进行训练或微调。

例如，样本语音是用户和智能对话系统之间的对话，在对话过程中，用户和智能对话系统可能谈论一些关键信息。例如，在订机票的场景下，智能对话系统可能会问用户需要哪天哪个时间的机票、哪种类型的机票（例如打折机票、非打折机票）等。而哪天哪个时间、哪种类型等均是关键信息。因此，本实施例可以通过该预训练模型预测出的关键信息和预先标注的关键信息，对该预训练模型进行训练或微调。

可以理解的是，根据情感分析任务、情绪识别任务、意图分析任务和/或关键信息收集任务对该预训练模型进行训练或微调时，可以在该预训练模型中部署新的全连接层，而不需要如上所述的全连接层 1、全连接层 2、全连接层 3 等。

本实施例首次在对话场景下，基于语音语义进行联合建模，根据语音的时序性特征，通过时序预测任务对第一机器学习模型进行预训练，并引入了跨模态对话回复选择任务对第一机器学习模型进行预训练。使得经过预训练后得到的预训练模型，不仅可以将语音信息和文本信息在时序上进行对齐，还可以识别上下文信息。经过实验证明，本实施例提供的方案比已有的语音语义建模方法有 5-20 个百分点的提升，能够更准确地理解用户的话语，提升用户体验。因此，在使用中，本实施例所述的方案不仅能学到更细粒度更准确的语音文本对齐，还可以捕捉到更多的语音信息，更好地帮助系统理解用户意图，提升用户体验。

另外，本实施例所述的方法比纯文本模型和纯语音模型都有所提升，其中比语音模型提升的更多一些，说明我们的模型很好地促成了语音语义的联合建模。并且，我们的方法相对已有最优的语音语义联合建模预训练方法提升是巨大的，即使是非预训练的方法，我们也在所有任务上都比各自的当前最先进的(state of the art, SOTA)方法提升了 1 个百分点以上，在多模态分析数据集（Multimodal Dataset, MOSI）上更是取得了 3.1 个百分点的提升。

此外，在公开的数据集上，与纯文本方法相比，本方案在准确率上可以实现 10.1% 的提升，即使与目前最先进的多模态方法相比，也可以实现 5-6% 的提升。

图 12 为本公开另一实施例提供的智能对话方法流程图。该方法可以由云端服务器来执行，例如，该云端服务器中部署有如上所述的经过训练或微调之后的机器学习模型，该云端服务器可以通过该训练或微调之后的机器学习模型来实现本实施例所述的智能对话方法。在本实施例中，该方法具体步骤如下：

S1201、获取目标用户和智能对话系统之间的历史对话，以及所述历史对话对应的文本信息。

例如，该云端服务器中可以预先部署有智能对话系统，该云端服务器可以获取目标用户和智能对话系统之间的历史对话，以及该历史对话对应的文本信息。

S1202、将所述历史对话和所述文本信息输入到训练后的机器学习模型中，使得所述机器学习模型输出所述目标用户的情感、情绪、意图、关键信息中的至少一种，所述机器学习模型是根据如上所述的方法训练得到的。

例如，该云端服务器可以将该历史对话和该文本信息输入到经过训练或微调之后的机器学习模型中，使得该机器学习模型输出该目标用户的情感、情绪、意图、关键

信息中的至少一种。

S1203、根据所述目标用户的情感、情绪、意图、关键信息中的至少一种，与所述目标用户进行智能对话。

例如，该云端服务器可以根据该目标用户的情感、情绪、意图、关键信息中的至少一种，与所述目标用户进行智能对话。例如，如果历史对话中的最后一句话是目标用户的提问，则该云端服务器可以根据该目标用户的情感、情绪、意图、关键信息中的至少一种，生成该提问的回复信息，并将该回复信息反馈给该目标用户的终端。如果历史对话中的最后一句话是智能对话系统的回复，则该云端服务器可以根据该目标用户的情感、情绪、意图、关键信息中的至少一种，生成问句，并将问句反馈给该目标用户的终端，从而使得该目标用户和该云端服务器中的智能对话系统继续进行对话。

本实施例通过获取目标用户和智能对话系统之间的历史对话，以及所述历史对话对应的文本信息，并将历史对话和所述文本信息输入到训练后的机器学习模型中，使得该机器学习模型输出所述目标用户的情感、情绪、意图、关键信息中的至少一种。从而提高了机器学习模型理解目标用户的精准度。进一步，根据所述目标用户的情感、情绪、意图、关键信息中的至少一种，与所述目标用户进行智能对话。使得该机器学习模型在精准理解目标用户的基础上，可以与目标用户进行精准的对话。

图 13 为本公开实施例提供的机器学习模型训练装置的结构示意图。本公开实施例提供的机器学习模型训练装置可以执行机器学习模型训练方法实施例提供的处理流程，如图 13 所示，机器学习模型训练装置 130 包括：

获取模块 131，被设置为获取语音信息、以及所述语音信息对应的文本信息；

编码模块 132，被设置为分别对所述文本信息和所述语音信息进行编码，得到文本编码信息和语音编码信息；

融合处理模块 133，被设置为对所述文本编码信息和所述语音编码信息进行融合处理，得到融合处理后的编码信息；

预测模块 134，被设置为根据所述融合处理后的编码信息中与所述文本信息对应的编码信息，预测所述文本信息中每个文本单元在所述语音信息中对应的时间信息；

预训练模块 135，被设置为根据所述预测的时间信息、以及所述文本信息中每个文本单元对应的语音单元在所述语音信息中的位置标注，对第一机器学习模型进行预训练。

可选的，编码模块 132 分别对所述文本信息和所述语音信息进行编码，得到文本编码信息和语音编码信息时，具体被设置为：

通过文本编码器对所述文本信息中每个文本单元分别包括的至少一个编码单元进行编码，得到文本编码信息，所述文本编码信息包括各个编码单元的表示向量；

通过语音编码器对所述语音信息中的每个语音片段进行编码，得到语音编码信息，所述语音编码信息包括各个语音片段的表示向量。

可选的，所述融合处理后的编码信息包括所述各个编码单元分别对应的融合表示向量、以及所述各个语音片段分别对应的融合表示向量；

预测模块 134 根据所述融合处理后的编码信息，预测所述文本信息中每个文本单元在所述语音信息中对应的时间信息时，具体被设置为：

若所述文本单元包括一个编码单元，则根据所述编码单元对应的融合表示向量，预测所述文本单元在所述语音信息中对应的开始时间和结束时间；

若所述文本单元包括多个编码单元，则根据所述多个编码单元中第一个编码单元对应的融合表示向量，预测所述文本单元在所述语音信息中对应的开始时间，以及根据所述多个编码单元中最后一个编码单元对应的融合表示向量，预测所述文本单元在所述语音信息中对应的结束时间。

可选的，预训练模块 135 根据所述预测的时间信息、以及所述文本信息中每个文

本单元对应的语音单元在所述语音信息中的位置标注,对第一机器学习模型进行预训练时,具体被设置为:

根据所述文本单元在所述语音信息中对应的开始时间和结束时间,以及预先标注的所述文本单元对应的语音单元在所述语音信息中的开始位置和结束位置,构建第一损失函数;

根据所述第一损失函数,对第一机器学习模型进行预训练。

可选的,所述第一损失函数包括第一差值和第二差值,所述第一差值是根据所述开始时间和所述开始位置得到的差值,所述第二差值是根据所述结束时间和所述结束位置得到的差值。

可选的,所述语音信息包括历史语音和当前语音,所述文本信息包括历史文本和当前文本;模型预训练装置 130 还包括:判断模块 136,被设置为根据所述融合处理后的编码信息,判断所述历史语音和所述当前语音是否匹配,得到第一判断结果,以及判断所述历史文本和所述当前文本是否匹配,得到第二判断结果;预训练模块 135 还被设置为根据所述第一判断结果和所述第二判断结果,预测输入样本的类型,所述输入样本包括所述语音信息和所述文本信息;根据所述预测的输入样本的类型、以及预先标注的所述输入样本的类型,构建第二损失函数,并根据所述第二损失函数对所述第一机器学习模型进行预训练,得到预训练模型。

图 13 所示实施例的模型预训练装置可被设置为执行上述方法实施例的技术方案,其实现原理和技术效果类似,此处不再赘述。

图 14 为本公开实施例提供的机器学习模型训练装置的结构示意图。本公开实施例提供的机器学习模型训练装置可以执行机器学习模型训练方法实施例提供的处理流程,如图 14 所示,机器学习模型训练装置 140 包括:

获取模块 141,被设置为获取目标任务的样本数据,所述样本数据包括待处理数据和所述待处理数据的标注结果,所述待处理数据包括样本语音、以及所述样本语音对应的样本文本;

输入模块 142,被设置为将所述待处理数据输入到第二机器学习模型中,使得所述第二机器学习模型输出处理后的结果,所述第二机器学习模型是根据如上所述的预训练方法得到的预训练模型;

训练模块 143,被设置为根据所述第二机器学习模型输出的处理后的结果、以及所述待处理数据的标注结果,对所述第二机器学习模型进行训练,得到训练后的机器学习模型。

可选的,所述目标任务包括如下至少一种:

情感分析任务、情绪识别任务、意图分析任务、关键信息收集任务。

图 14 所示实施例的模型训练装置可被设置为执行上述方法实施例的技术方案,其实现原理和技术效果类似,此处不再赘述。

图 15 为本公开实施例提供的智能对话装置的结构示意图。本公开实施例提供的智能对话装置可以执行智能对话方法实施例提供的处理流程,如图 15 所示,智能对话装置 150 包括:

获取模块 151,被设置为获取目标用户和智能对话系统之间的历史对话,以及所述历史对话对应的文本信息;

输入模块 152,被设置为将所述历史对话和所述文本信息输入到训练后的机器学习模型中,使得所述机器学习模型输出所述目标用户的情感、情绪、意图、关键信息中的至少一种,所述机器学习模型是根据如上所述的方法训练得到的;

对话模块 153,被设置为根据所述目标用户的情感、情绪、意图、关键信息中的至少一种,与所述目标用户进行智能对话。

图 15 所示实施例的智能对话装置可被设置为执行上述方法实施例的技术方案,

其实现原理和技术效果类似，此处不再赘述。

以上描述了模型预训练装置、的内部功能和结构，该装置可实现为一种电子设备。图 16 为本公开实施例提供的电子设备实施例的结构示意图。如图 16 所示，该电子设备包括存储器 161 和处理器 162。

存储器 161 被设置为存储程序。除上述程序之外，存储器 161 还可被配置为存储其它各种数据以支持在电子设备上的操作。这些数据的示例包括被设置为在电子设备上操作的任何应用程序或方法的指令，联系人数据，电话簿数据，消息，图片，视频等。

存储器 161 可以由任何类型的易失性或非易失性存储设备或者它们的组合实现，如静态随机存取存储器(SRAM)，电可擦除可编程只读存储器(EEPROM)，可擦除可编程只读存储器(EPROM)，可编程只读存储器(PROM)，只读存储器(ROM)，磁存储器，闪存存储器，磁盘或光盘。

处理器 162 与存储器 161 耦合，执行存储器 161 所存储的程序，以被设置为：

获取语音信息、以及所述语音信息对应的文本信息；

分别对所述文本信息和所述语音信息进行编码，得到文本编码信息和语音编码信息；

对所述文本编码信息和所述语音编码信息进行融合处理，得到融合处理后的编码信息；

根据所述融合处理后的编码信息中与所述文本信息对应的编码信息，预测所述文本信息中每个文本单元在所述语音信息中对应的时间信息；

根据所述预测的时间信息、以及所述文本信息中每个文本单元对应的语音单元在所述语音信息中的位置标注，对第一机器学习模型进行预训练，得到预训练模型。

进一步，如图 16 所示，电子设备还可以包括：通信组件 163、电源组件 164、音频组件 165、显示器 166 等其它组件。图 16 中仅示意性给出部分组件，并不意味着电子设备只包括图 16 所示组件。

通信组件 163 被配置为便于电子设备和其他设备之间有线或无线方式的通信。电子设备可以接入基于通信标准的无线网络，如 WiFi，2G 或 3G，或它们的组合。在一个示例性实施例中，通信组件 163 经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中，所述通信组件 163 还包括近场通信(NFC)模块，以促进短程通信。例如，在 NFC 模块可基于射频识别(RFID)技术，红外数据协会(IrDA)技术，超宽带(UWB)技术，蓝牙(BT)技术和其他技术来实现。

电源组件 164，为电子设备的各种组件提供电力。电源组件 164 可以包括电源管理系统，一个或多个电源，及其他与为电子设备生成、管理和分配电力相关联的组件。

音频组件 165 被配置为输出和/或输入音频信号。例如，音频组件 165 包括一个麦克风(MIC)，当电子设备处于操作模式，如呼叫模式、记录模式和语音识别模式时，麦克风被配置为接收外部音频信号。所接收的音频信号可以被进一步存储在存储器 161 或经由通信组件 163 发送。在一些实施例中，音频组件 165 还包括一个扬声器，被设置为输出音频信号。

显示器 166 包括屏幕，其屏幕可以包括液晶显示器(LCD)和触摸面板(TP)。如果屏幕包括触摸面板，屏幕可以被实现为触摸屏，以接收来自用户的输入信号。触摸面板包括一个或多个触摸传感器以感测触摸、滑动和触摸面板上的手势。所述触摸传感器可以不仅感测触摸或滑动动作的边界，而且还检测与所述触摸或滑动操作相关的持续时间和压力。

另外，本公开实施例还提供一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行以实现上述实施例所述的模型预训练、训练、智能对话方法。

需要说明的是，在本文中，诸如“第一”和“第二”等之类的关系术语用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

以上所述是本公开的具体实施方式，使本领域技术人员能够理解或实现本公开。对这些实施例的多种修改对本领域的技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本公开的精神或范围的情况下，在其它实施例中实现。因此，本公开将不会被限制于本文所述的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

工业实用性

本公开实施例所提供的机器学习模型训练、对话方法、装置、设备及介质应用于机器学习场景中，通过获取语音信息、以及所述语音信息对应的文本信息，并分别对所述文本信息和所述语音信息进行编码，得到文本编码信息和语音编码信息。进一步，对所述文本编码信息和所述语音编码信息进行融合处理，得到融合处理后的编码信息，根据所述融合处理后的编码信息，预测所述文本信息中每个文本单元在所述语音信息中对应的时间信息，然后根据所述预测的时间信息、以及所述文本信息中每个文本单元对应的语音单元在所述语音信息中的位置标注，对第一机器学习模型进行预训练，得到预训练模型。使得预训练模型可以将语音信息和文本信息在时序上进行对齐，例如，文本信息中更细粒度的文本单元与语音信息中更细粒度的语音单元之间的对齐，从而提高预训练模型对语音信息和文本信息的理解力，从而提高预训练模型对用户话语理解的精准度。

权利要求书

1. 一种机器学习模型训练方法，所述方法包括：
 - 获取语音信息、以及所述语音信息对应的文本信息；
 - 分别对所述文本信息和所述语音信息进行编码，得到文本编码信息和语音编码信息；
 - 对所述文本编码信息和所述语音编码信息进行融合处理，得到融合处理后的编码信息；
 - 根据所述融合处理后的编码信息中与所述文本信息对应的编码信息，预测所述文本信息中每个文本单元在所述语音信息中对应的时间信息；
 - 根据所述预测的时间信息、以及所述文本信息中每个文本单元对应的语音单元在所述语音信息中的位置标注，对第一机器学习模型进行预训练，得到预训练模型。
2. 根据权利要求1所述的方法，其中，分别对所述文本信息和所述语音信息进行编码，得到文本编码信息和语音编码信息，包括：
 - 通过文本编码器对所述文本信息中每个文本单元分别包括的至少一个编码单元进行编码，得到文本编码信息，所述文本编码信息包括各个编码单元的代表向量；
 - 通过语音编码器对所述语音信息中的每个语音片段进行编码，得到语音编码信息，所述语音编码信息包括各个语音片段的代表向量。
3. 根据权利要求2所述的方法，其中，所述融合处理后的编码信息包括所述各个编码单元分别对应的融合表示向量、以及所述各个语音片段分别对应的融合表示向量；
 - 根据所述融合处理后的编码信息，预测所述文本信息中每个文本单元在所述语音信息中对应的时间信息，包括：
 - 若所述文本单元包括一个编码单元，则根据所述编码单元对应的融合表示向量，预测所述文本单元在所述语音信息中对应的开始时间和结束时间；
 - 若所述文本单元包括多个编码单元，则根据所述多个编码单元中第一个编码单元对应的融合表示向量，预测所述文本单元在所述语音信息中对应的开始时间，以及根据所述多个编码单元中最后一个编码单元对应的融合表示向量，预测所述文本单元在所述语音信息中对应的结束时间。
4. 根据权利要求3所述的方法，其中，根据所述预测的时间信息、以及所述文本信息中每个文本单元对应的语音单元在所述语音信息中的位置标注，对第一机器学习模型进行预训练，包括：
 - 根据所述文本单元在所述语音信息中对应的开始时间和结束时间，以及预先标注的所述文本单元对应的语音单元在所述语音信息中的开始位置和结束位置，构建第一损失函数；
 - 根据所述第一损失函数，对第一机器学习模型进行预训练。
5. 根据权利要求4所述的方法，其中，所述第一损失函数包括第一差值和第二差值，所述第一差值是根据所述开始时间和所述开始位置得到的差值，所述第二差值是根据所述结束时间和所述结束位置得到的差值。
6. 根据权利要求1所述的方法，其中，所述语音信息包括历史语音和当前语音，所述文本信息包括历史文本和当前文本；
 - 所述方法还包括：
 - 根据所述融合处理后的编码信息，判断所述历史语音和所述当前语音是否匹配，得到第一判断结果，以及判断所述历史文本和所述当前文本是否匹配，得到第二判断结果；
 - 根据所述第一判断结果和所述第二判断结果，预测输入样本的类型，所述输入样本包括所述语音信息和所述文本信息；

根据所述预测的输入样本的类型、以及预先标注的所述输入样本的类型，构建第二损失函数，并根据所述第二损失函数对所述第一机器学习模型进行预训练，得到预训练模型。

7. 一种机器学习模型训练方法，所述方法包括：

获取目标任务的样本数据，所述样本数据包括待处理数据和所述待处理数据的标注结果，所述待处理数据包括样本语音、以及所述样本语音对应的样本文本；

将所述待处理数据输入到第二机器学习模型中，使得所述第二机器学习模型输出处理后的结果，所述第二机器学习模型是根据如权利要求 1-6 任一项所述的预训练方法得到的预训练模型；

根据所述第二机器学习模型输出的处理后的结果、以及所述待处理数据的标注结果，对所述第二机器学习模型进行训练，得到训练后的机器学习模型。

8. 根据权利要求 7 所述的方法，其中，所述目标任务包括如下至少一种：

情感分析任务、情绪识别任务、意图分析任务、关键信息收集任务。

9. 一种智能对话方法，所述方法包括：

获取目标用户和智能对话系统之间的历史对话，以及所述历史对话对应的文本信息；

将所述历史对话和所述文本信息输入到训练后的机器学习模型中，使得所述机器学习模型输出所述目标用户的情感、情绪、意图、关键信息中的至少一种，所述机器学习模型是根据权利要求 7 所述的方法训练得到的；

根据所述目标用户的情感、情绪、意图、关键信息中的至少一种，与所述目标用户进行智能对话。

10. 一种机器学习模型训练装置，包括：

获取模块，被设置为获取语音信息、以及所述语音信息对应的文本信息；

编码模块，被设置为分别对所述文本信息和所述语音信息进行编码，得到文本编码信息和语音编码信息；

融合处理模块，被设置为对所述文本编码信息和所述语音编码信息进行融合处理，得到融合处理后的编码信息；

预测模块，被设置为根据所述融合处理后的编码信息中与所述文本信息对应的编码信息，预测所述文本信息中每个文本单元在所述语音信息中对应的时间信息；

预训练模块，被设置为根据所述预测的时间信息、以及所述文本信息中每个文本单元对应的语音单元在所述语音信息中的位置标注，对第一机器学习模型进行预训练。

11. 一种电子设备，包括：

存储器；

处理器；以及

计算机程序；

其中，所述计算机程序存储在所述存储器中，并被配置为由所述处理器执行以实现如权利要求 1-9 中任一项所述的方法。

12. 一种计算机可读存储介质，其上存储有计算机程序，其中，所述计算机程序被处理器执行时实现如权利要求 1-9 中任一项所述的方法。

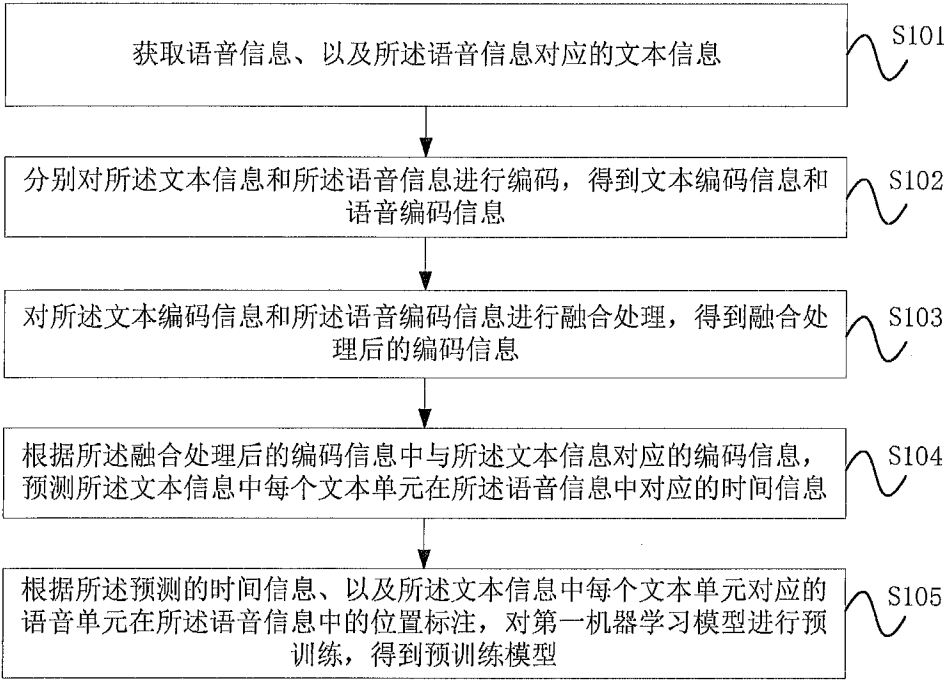


图 1

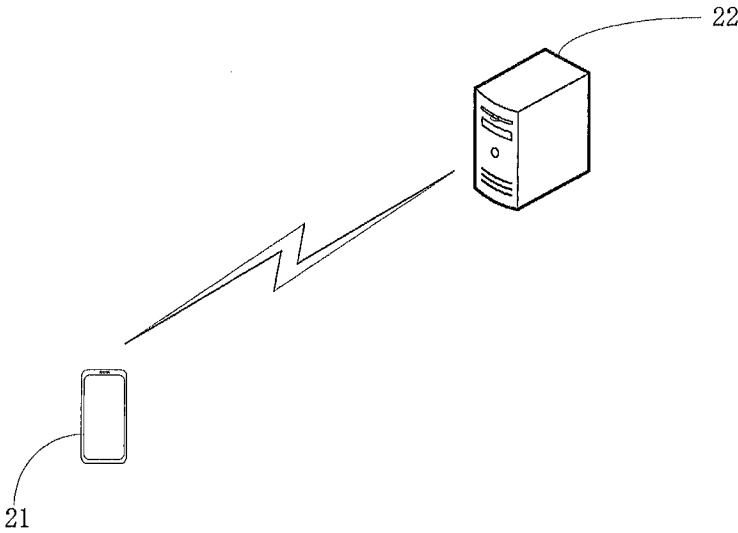


图 2

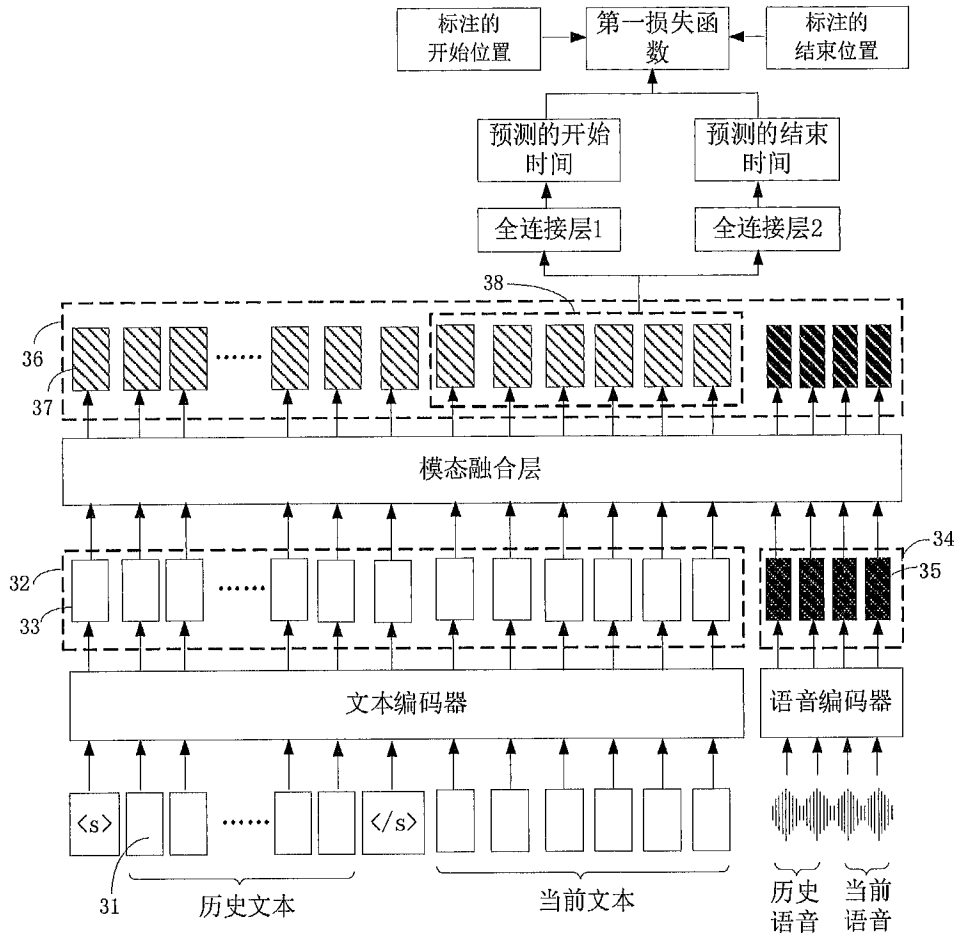


图 3



图 4

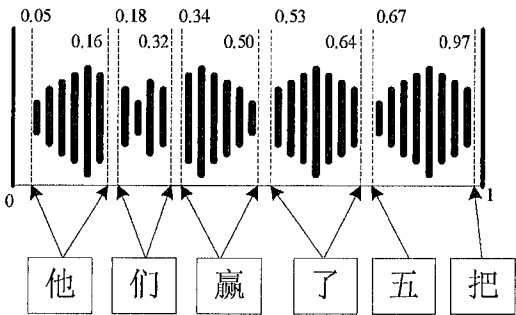


图 5

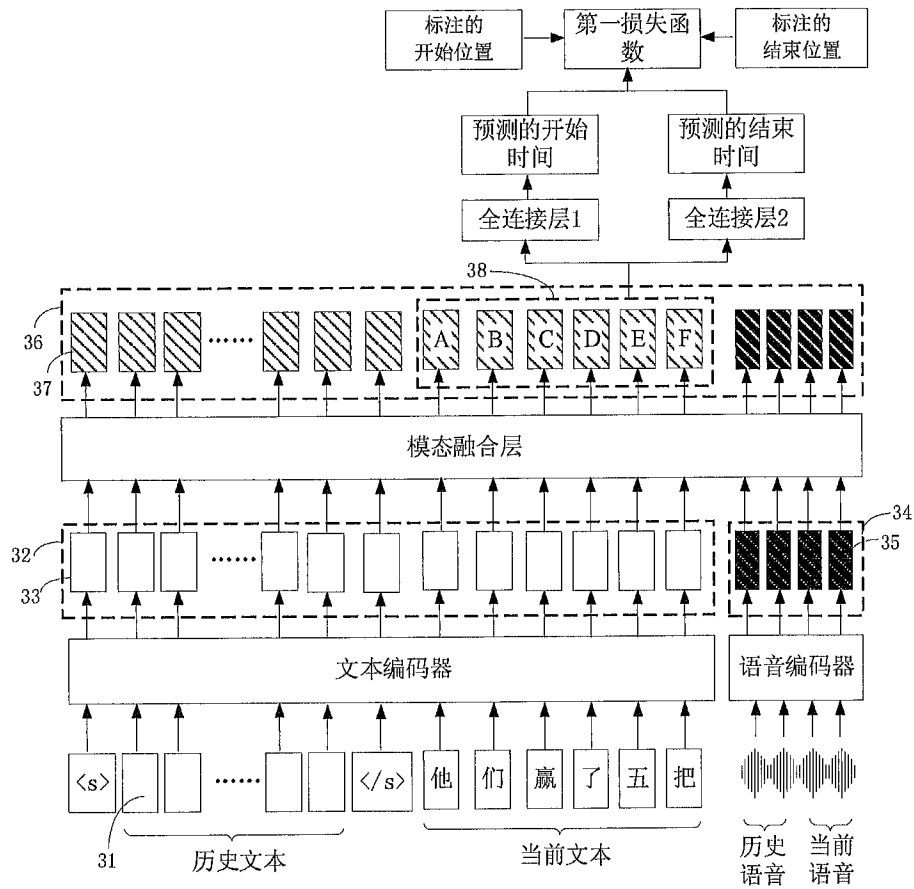


图 6

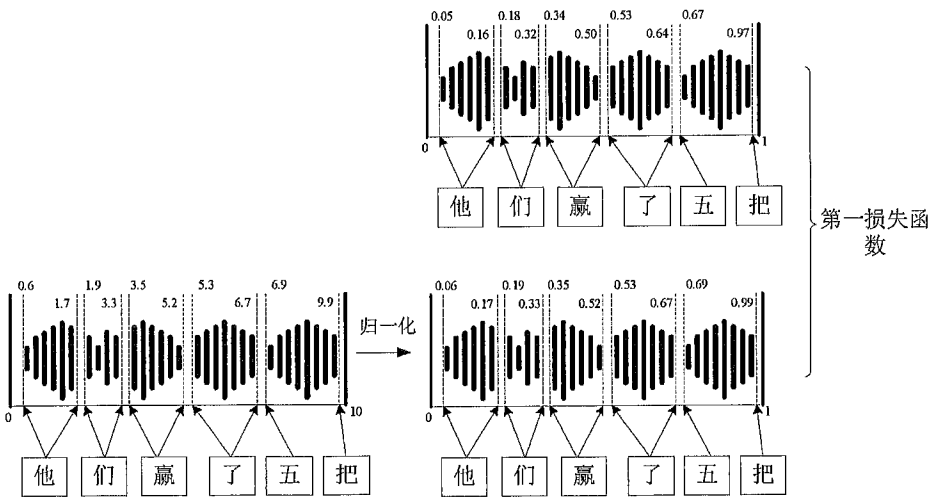


图 7

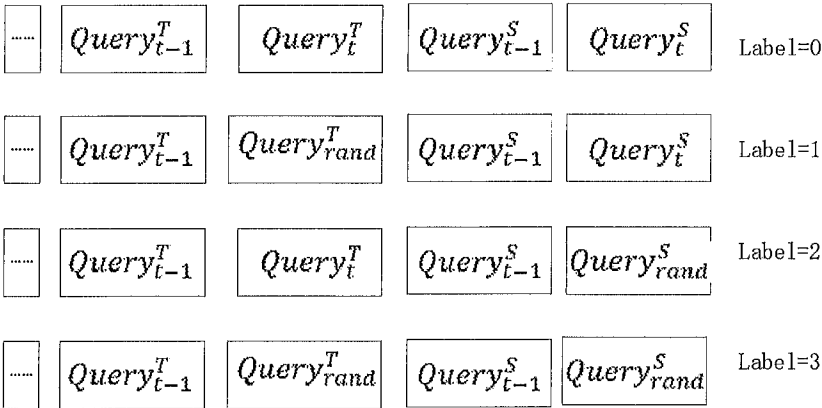


图 8

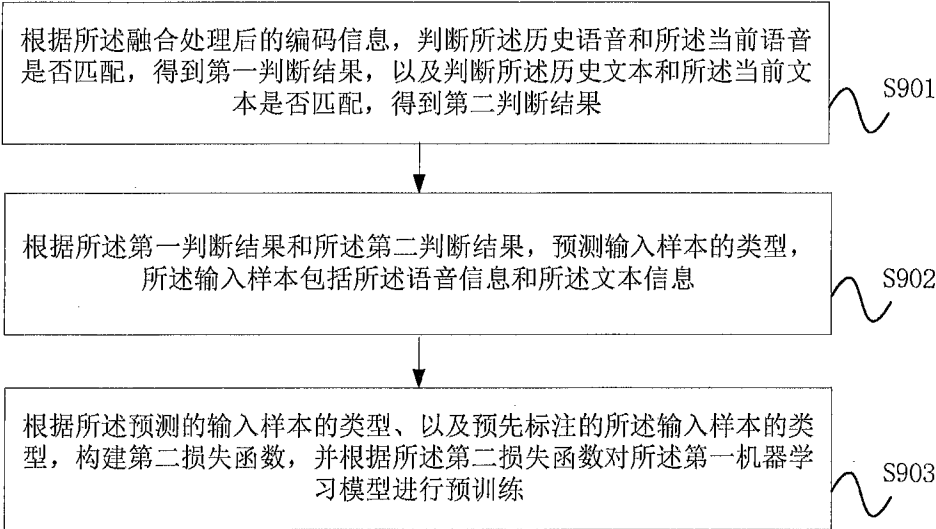


图 9

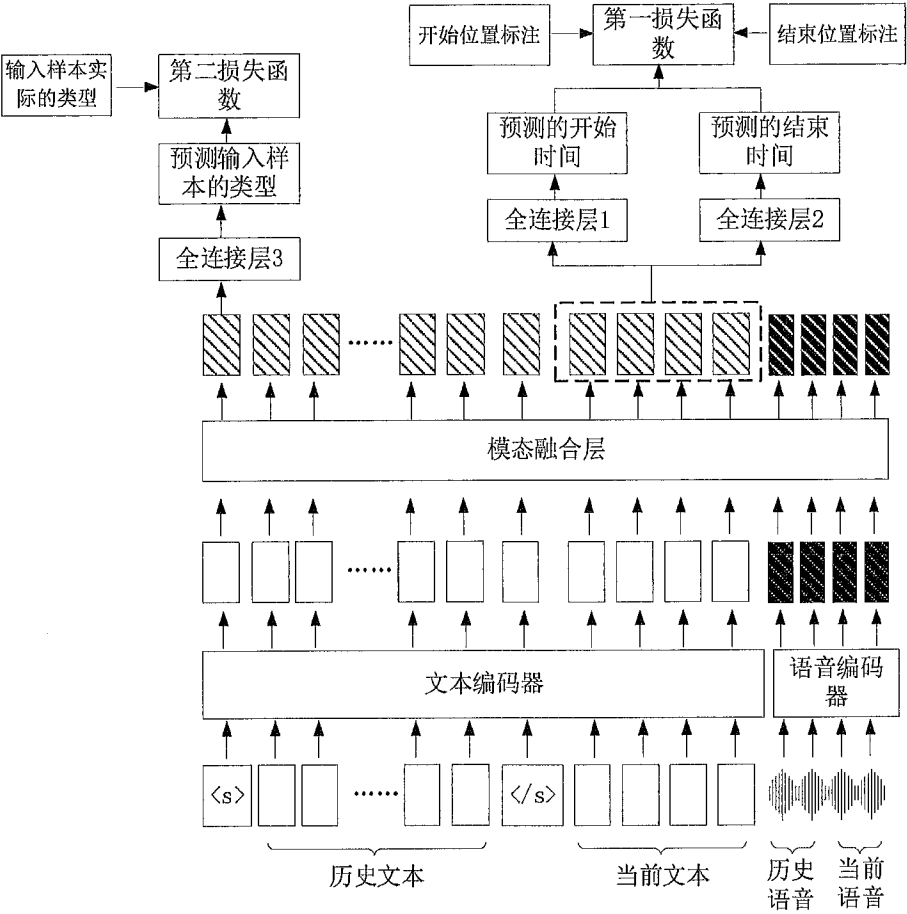


图 10

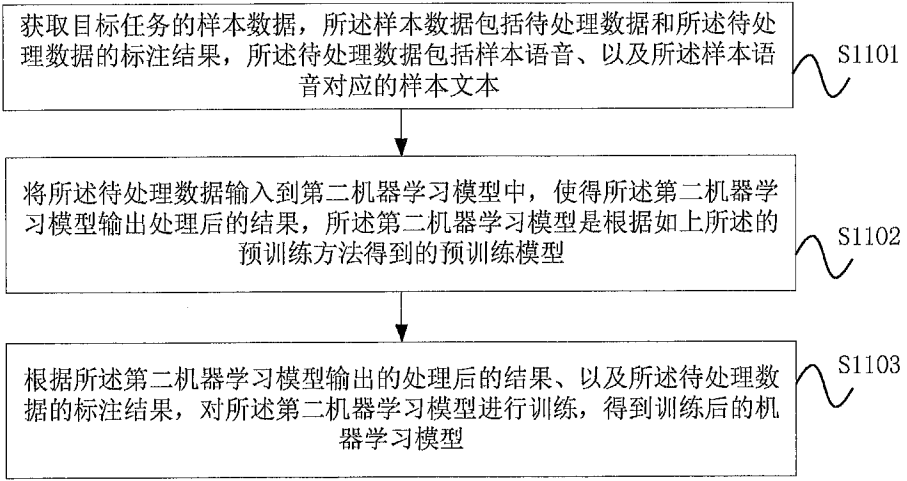


图 11

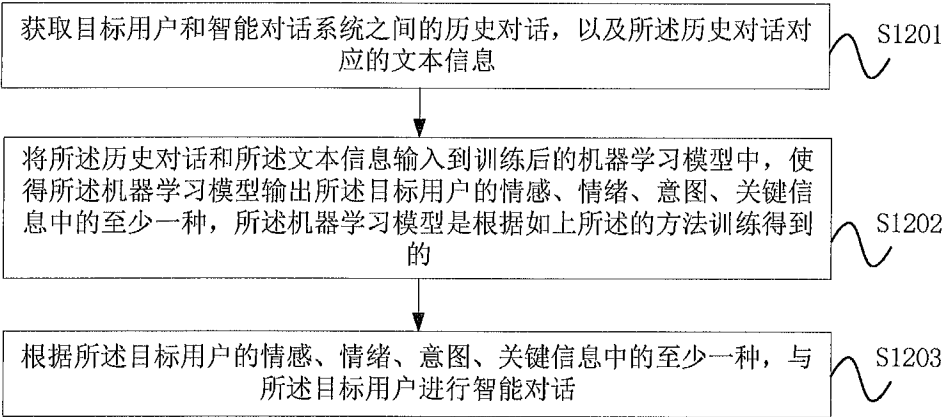


图 12

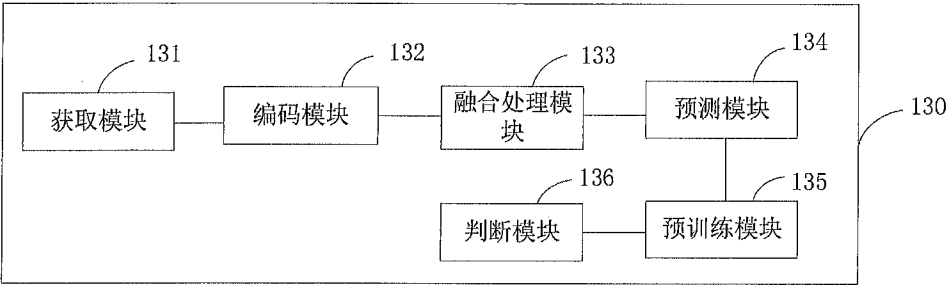


图 13

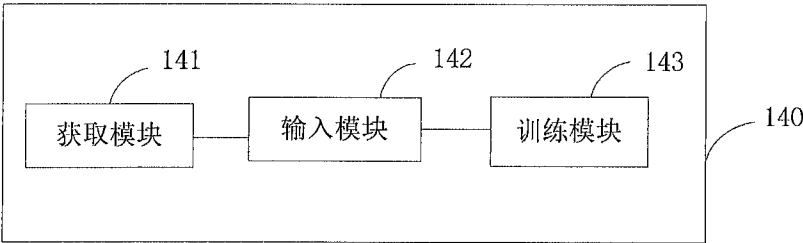


图 14

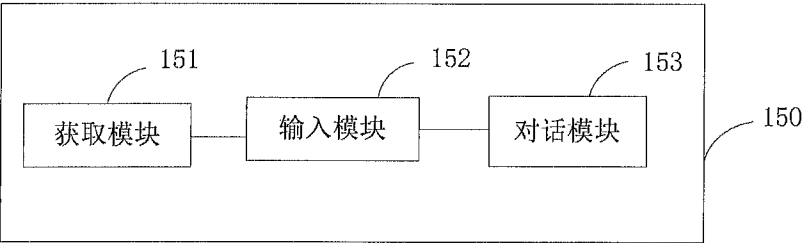


图 15

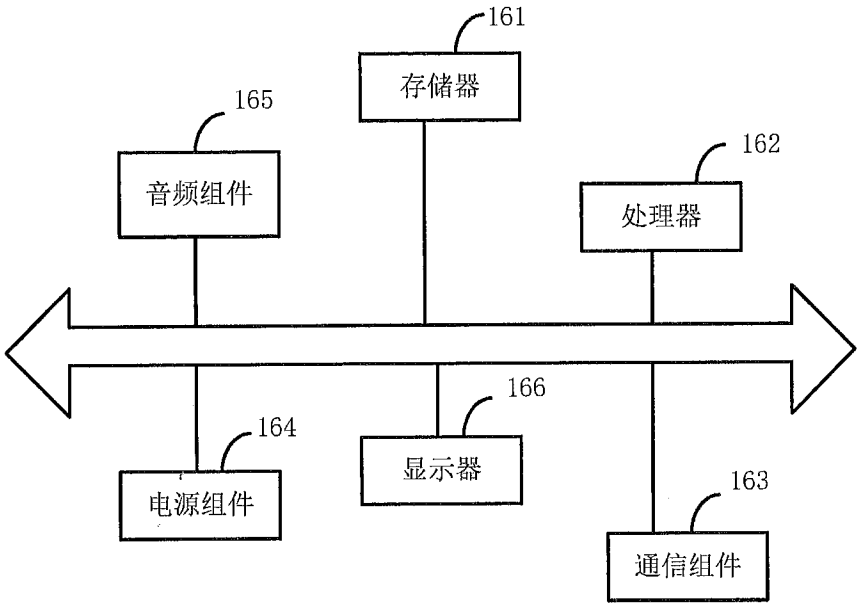


图 16

INTERNATIONAL SEARCH REPORT

International application No.

PCT/SG2024/050199

A. CLASSIFICATION OF SUBJECT MATTER

See Extra Box

According to International Patent Classification (IPC)

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L, G06F, G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

FAMPAT, IEEE, CNKI: 预测, 时间, 文本, 语音, 标注, 机器学习, predict, time, text, audio, label, machine learning and other related search terms.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 114360504 A (TENCENT SCIENCE AND TECHNOLOGY (SHENZHEN) LIMITED COMPANY), 15 April 2022, figures 3A and 3C, and paragraphs [0027]-[0030], [0075], [0078], [0080], [0081], [0083], [0084], [0098], [0101]-[0104], [0106]-[0108], [0110] and [0120]-[0123]	1-12
A	CN 114267345 A (ALIBABA DHARMA ACADEMY (HANGZHOU) TECHNOLOGY CO., LTD.), 01 April 2022, entire document	

☒ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
“A” document defining the general state of the art which is not considered to be of particular relevance	“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
“D” document cited by the applicant in the international application	“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
“E” earlier application or patent but published on or after the international filing date	“&” document member of the same patent family
“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	
“O” document referring to an oral disclosure, use, exhibition or other means	
“P” document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 25 June 2024	Date of the issuance of the international search report 25 June 2024
Name and mailing address of the ISA/SG  Intellectual Property Office of Singapore 1 Paya Lebar Link, #11-03 PLQ 1, Paya Lebar Quarter Singapore 408533 E-mail: pct@ipos.gov.sg	Authorized officer ZHAO, Xue (Dr.) IPOS Telephone No. (+65) 6339 8616

INTERNATIONAL SEARCH REPORT

International application No.

PCT/SG2024/050199

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 114691852 A (ALIBABA DHARMA ACADEMY (HANGZHOU) TECHNOLOGY CO., LTD.), 01 July 2022, entire document	
A	CN 114936560 A (ALIBABA (CHINA) CO., LTD.), 23 August 2022, entire document	

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/SG2024/050199

Note: related known patent family members of the patent documents referred to by this international search report are listed in this annex. This authority only provides information on the related patent family members, and assumes no responsibility for the details of the patents.

Patent Documents referred in the Report	Publication Date/Announcement Date (day/month/year)	Patent Family	Publication Date/Announcement Date (day/month/year)
CN 114360504 A	15/04/2022	US 2023/0306959 A1 WO 2023/093295 A1	28/09/2023 01/06/2023
CN 114267345 A	01/04/2022	None	
CN 114691852 A	01/07/2022	US 2023/0395075 A1	07/12/2023
CN 114936560 A	23/08/2022	None	

INTERNATIONAL SEARCH REPORT

International application No.

PCT/SG2024/050199

EXTRA BOX

(CLASSIFICATION OF SUBJECT MATTER)

International Classification Symbol:

G10L 15/00 (2013.01)

G06F 40/00 (2020.01)

G06N 20/00 (2019.01)

G06N 3/08 (2023.01)

G06F 16/332 (2019.01)

国际检索报告

国际申请号

PCT/SG2024/050199

A. 主题的分类

请参照附加栏

按照国际专利分类(IPC)

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G10L, G06F, G06N

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

FAMPAT, IEEE, CNKI: 预测, 时间, 文本, 语音, 标注, 机器学习, predict, time, text, audio, label, machine learning 及其他相关检索词。

C. 相关文件

类型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 114360504 A (腾讯科技 (深圳) 有限公司) 2022 年 04 月 15 日 图 3A, 3C, 第[0027]-[0030], [0075], [0078], [0080], [0081], [0083], [0084], [0098], [0101]-[0104], [0106]-[0108], [0110], [0120]-[0123]段	1-12
A	CN 114267345 A (阿里巴巴达摩院 (杭州) 科技有限公司) 2022 年 04 月 01 日 全文	

☒ 其余文件在 C 栏的续页中列出。

☒ 见同族专利附件。

* 引用文件的具体类型:

"A" 认为不特别相关的表示了现有技术一般状态的文件

"D" 在此国际申请中引用的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期

2024 年 06 月 25 日

国际检索报告发文日

2024 年 06 月 25 日

ISA/SG 的名称和邮寄地址

新加坡知识产权局



巴耶利巴连路 1 号 #11-03

PLQ 1, 巴耶利峇中心

新加坡 408533

电子邮件账号: pct@ipos.gov.sg

受权官员

赵雪(博士)

IPOS 客服电话号码: (+65) 6339 8616

国际检索报告

国际申请号

PCT/SG2024/050199

C (续). 相关文件

类型*	引用文件, 必要时, 指明相关段落	相关的权利要求
A	CN 114691852 A (阿里巴巴达摩院 (杭州) 科技有限公司) 2022 年 07 月 01 日 全文	
A	CN 114936560 A (阿里巴巴 (中国) 有限公司) 2022 年 08 月 23 日 全文	

国际检索报告

关于同族专利的信息

国际申请号

PCT/SG2024/050199

备注: 此附件列出了本国际检索报告引用的专利文件的相关已知同族专利。本单位仅提供相关同族专利的信息, 并不对专利的详情承担任何责任。

检索报告中引用的专利文件	公布日或公告日	同族专利	公布日或公告日
CN 114360504 A	15/04/2022	US 2023/0306959 A1 WO 2023/093295 A1	28/09/2023 01/06/2023
CN 114267345 A	01/04/2022	无	
CN 114691852 A	01/07/2022	US 2023/0395075 A1	07/12/2023
CN 114936560 A	23/08/2022	无	

附加栏
(主题的分类)

国际分类号:
G10L 15/00 (2013.01)
G06F 40/00 (2020.01)
G06N 20/00 (2019.01)
G06N 3/08 (2023.01)
G06F 16/332 (2019.01)