(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107783960 B
(45)授权公告日 2021.07.23

(21)申请号 201710996056.0

(22)申请日 2017.10.23

(65)同一申请的已公布的文献号
申请公布号 CN 107783960 A

(43)申请公布日 2018.03.09

(73)专利权人 百度在线网络技术(北京)有限公司
地址 100085 北京市海淀区上地十街10号百度大厦三层

(72)发明人 孙明明 李平

(74)专利代理机构 北京英赛嘉华知识产权代理有限责任公司 11204
代理人 王达佐 马晓亚

(51)Int.Cl.
G06F 40/253(2020.01)
G06F 40/284(2020.01)
G06F 40/211(2020.01)
G06F 40/30(2020.01)
G06F 16/28(2019.01)
G06N 3/08(2006.01)

审查员 周林

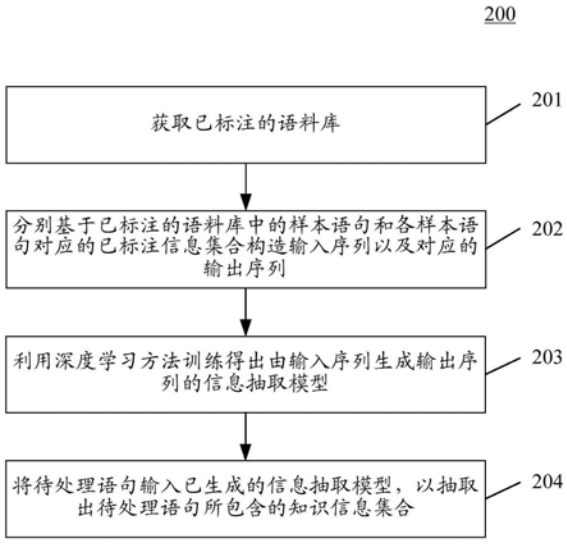
权利要求书4页 说明书17页 附图4页

(54)发明名称

用于抽取信息的方法、装置和设备

(57)摘要

本申请公开了用于抽取信息的方法、装置和设备。用于抽取信息的方法的一具体实施方式包括:获取已标注的语料库,已标注的语料库包括多条样本语句以及与各样本语句对应的已标注信息集合;基于已标注的语料库中的样本语句和各样本语句对应的已标注信息集合构造输入序列和输出序列;利用深度学习方法训练得出由输入序列生成输出序列的信息抽取模型;将待处理语句输入信息抽取模型,以抽取出待处理语句所包含的知识信息集合;已标注信息集合包括将从对应的样本语句中抽取出的以下至少一种知识按照预设规则规范化后的信息:基于动词或介词的知识、基于名词属性的知识、实体的描述知识以及实体与概念的关系知识。该实施方式实现了信息的有效、准确抽取。



CN 107783960 B

1.一种用于抽取信息的方法,其特征在于,所述方法包括:

获取已标注的语料库,所述已标注的语料库包括多条样本语句以及与各样本语句对应的已标注信息集合;

分别基于所述已标注的语料库中的样本语句和各所述样本语句对应的已标注信息集合构造输入序列以及对应的输出序列;

利用深度学习方法训练得出由所述输入序列生成所述输出序列的信息抽取模型;

将待处理语句输入已生成的信息抽取模型,以抽取出所述待处理语句所包含的知识信息集合;

其中,所述已标注信息集合包括将从对应的样本语句中抽取出的以下至少一种知识按照预设规则规范化后的信息:

基于动词或介词的知识、基于名词属性的知识、实体的描述知识以及实体与概念的关系知识;其中,所述预设规则是预先定义的采用符号协助的对样本语句中的各类型知识的表达方式。

2.根据权利要求1所述的方法,其特征在于,所述利用深度学习方法训练得出由所述输入序列生成所述输出序列的信息抽取模型,包括:

对所述输入序列中的各对象和所述输出序列中的各已标注信息进行向量化处理;

构建所述信息抽取模型的神经网络,所述信息抽取模型的神经网络包括编码器和解码器;

利用所述编码器将向量化处理后的输入序列映射为输入隐状态序列,利用所述解码器将所述输出序列映射为输出隐状态序列;

利用解码器对所述输入隐状态序列进行解码,得到预测状态序列;

根据所述预测状态序列与所述输出隐状态序列的差异调整所述神经网络的参数。

3.根据权利要求2所述的方法,其特征在于,所述利用解码器对所述输入隐状态序列进行解码,得到预测状态序列,包括:

对于待生成的预测状态序列中的目标位置,获取所述解码器的隐藏层在解码得出所述目标位置的上一个位置的预测状态后的状态,作为所述解码器的当前隐状态;

基于所述当前隐状态计算所述输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度;

基于匹配度计算各所述输入隐状态对所述目标位置的预测状态的注意力权重;

对各所述输入隐状态按照所述注意力权重进行加权求和得到上下文向量;

基于所述上下文向量、所述输出隐状态序列中目标位置的上一个位置的输出隐状态、以及所述解码器的隐藏层在解码所述目标位置的预测状态时的状态,计算出所述目标位置的预测状态的概率分布;

根据所述概率分布,确定出所述目标位置的预测状态。

4.根据权利要求3所述的方法,其特征在于,所述基于所述当前隐状态计算所述输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度,包括:

对所述当前隐状态和所述输入隐状态进行第一非线性变换,得到所述输入隐状态与所述目标位置的预测状态之间的匹配度。

5.根据权利要求3所述的方法，其特征在于，所述基于所述当前隐状态计算所述输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度，包括：

基于所述当前隐状态计算所述目标位置对应的覆盖向量，所述覆盖向量用于记录所述样本语句中已被提取的信息；

对所述当前隐状态、所述输入隐状态以及所述目标位置的上一个位置对应的覆盖向量进行第二非线性变换，得到所述输入隐状态与所述目标位置的预测状态之间的匹配度。

6.根据权利要求3或4所述的方法，其特征在于，在基于所述当前隐状态计算所述输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度之前，所述利用解码器对所述输入隐状态序列进行解码，得到预测状态序列，还包括：

采用自然语言处理方法提取所述样本语句中的语义依存关系树；

对所述语义依存关系树中的每一条边添加一条反向边，构成路径查找图；

在路径查找图中查找所述输入序列中任意两个对象之间的语义路径，并将该语义路径编码为语义向量；

对所述语义向量进行变换得到语义依存关系注意力向量；

所述基于所述当前隐状态计算所述输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度，包括：

基于所述当前隐状态、所述输入隐状态以及所述语义依存关系注意力向量进行第三非线性变换，得到所述输入隐状态与所述目标位置的预测状态之间的匹配度。

7.根据权利要求3所述的方法，其特征在于，所述目标位置的预测状态的概率为：从对应的样本语句中拷贝一个词语作为所述输出序列中的一个对象的概率与从预设符号集合中选择出一个目标符号并将目标符号所表征的对象作为所述输出序列中的一个对象的概率之和；

所述预设符号集合中的符号用于与所述样本语句中的词语联合，以完整表征所述样本语句中的以下知识中的一项：基于动词或介词的知识、基于名词属性的知识、实体的描述知识以及实体与概念的关系知识。

8.根据权利要求7所述的方法，其特征在于，所述预设符号集合中的符号包括：

用于表征语句中分离的谓词的符号；

用于表征语句中以缩减表达式表达的知识中被缩减的信息的符号；

用于表征语句中的隐含知识的符号；

用于表征语句中的缺失信息的符号。

9.一种用于抽取信息的装置，其特征在于，所述装置包括：

获取单元，用于获取已标注的语料库，所述已标注的语料库包括多条样本语句以及与各样本语句对应的已标注信息集合；

构造单元，用于分别基于所述已标注的语料库中的样本语句和各所述样本语句对应的已标注信息集合构造输入序列以及对应的输出序列；

训练单元，用于利用深度学习装置训练得出由所述输入序列生成所述输出序列的信息抽取模型；

抽取单元，用于将待处理语句输入已生成的信息抽取模型，以抽取出所述待处理语句

所包含的知识信息集合；

其中，所述已标注信息集合包括将从对应的样本语句中抽取出的以下至少一种知识按照预设规则规范化后的信息：

基于动词或介词的知识、基于名词属性的知识、实体的描述知识以及实体与概念的关系知识；其中，所述预设规则是预先定义的采用符号协助的对样本语句中的各类型知识的表达方式。

10.根据权利要求9所述的装置，其特征在于，所述训练单元进一步用于按照如下方式训练得出由所述输入序列生成所述输出序列的信息抽取模型：

对所述输入序列中的各对象和所述输出序列中的各已标注信息进行向量化处理；

构建所述信息抽取模型的神经网络，所述信息抽取模型的神经网络包括编码器和解码器；

利用所述编码器将向量化处理后的输入序列映射为输入隐状态序列，利用所述解码器将所述输出序列映射为输出隐状态序列；

利用解码器对所述输入隐状态序列进行解码，得到预测状态序列；

根据所述预测状态序列与所述输出隐状态序列的差异调整所述神经网络的参数。

11.根据权利要求10所述的装置，其特征在于，所述训练单元进一步用于按照如下方式得到预测状态序列：

对于待生成的预测状态序列中的目标位置，获取所述解码器的隐藏层在解码得出所述目标位置的上一个位置的预测状态后的状态，作为所述解码器的当前隐状态；

基于所述当前隐状态计算所述输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度；

基于匹配度计算各所述输入隐状态对所述目标位置的预测状态的注意力权重；

对各所述输入隐状态按照所述注意力权重进行加权求和得到上下文向量；

基于所述上下文向量、所述输出隐状态序列中目标位置的上一个位置的输出隐状态、以及所述解码器的隐藏层在解码所述目标位置的预测状态时的状态，计算出所述目标位置的预测状态的概率分布；

根据所述概率分布，确定出所述目标位置的预测状态。

12.根据权利要求11所述的装置，其特征在于，所述训练单元进一步用于按照如下方式计算所述输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度：

对所述当前隐状态和所述输入隐状态进行第一非线性变换，得到所述输入隐状态与所述目标位置的预测状态之间的匹配度。

13.根据权利要求11所述的装置，其特征在于，所述训练单元进一步用于按照如下方式计算所述输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度：

基于所述当前隐状态计算所述目标位置对应的覆盖向量，所述覆盖向量用于记录所述样本语句中已被提取的信息；

对所述当前隐状态、所述输入隐状态以及所述目标位置的上一个位置对应的覆盖向量进行第二非线性变换，得到所述输入隐状态与所述目标位置的预测状态之间的匹配度。

14.根据权利要求11或12所述的装置，其特征在于，所述训练单元还用于在基于所述当前隐状态计算所述输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度之前执行如下操作：

采用自然语言处理装置提取所述样本语句中的语义依存关系树；

对所述语义依存关系树中的每一条边添加一条反向边，构成路径查找图；

在路径查找图中查找所述输入序列中任意两个对象之间的语义路径，并将该语义路径编码为语义向量；

对所述语义向量进行变换得到语义依存关系注意力向量；

所述训练单元用于按照如下方式计算所述输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度：

基于所述当前隐状态、所述输入隐状态以及所述语义依存关系注意力向量进行第三非线性变换，得到所述输入隐状态与所述目标位置的预测状态之间的匹配度。

15.根据权利要求11所述的装置，其特征在于，所述目标位置的预测状态的概率为：从对应的样本语句中拷贝一个词语作为所述输出序列中的一个对象的概率与从预设符号集合中选择出一个目标符号并将目标符号所表征的对象作为所述输出序列中的一个对象的概率之和；

所述预设符号集合中的符号用于与所述样本语句中的词语联合，以完整表征所述样本语句中的以下知识中的一项：基于动词或介词的知识、基于名词属性的知识、实体的描述知识以及实体与概念的关系知识。

16.根据权利要求15所述的装置，其特征在于，所述预设符号集合中的符号包括：

用于表征语句中分离的谓词的符号；

用于表征语句中以缩减表达式表达的知识中被缩减的信息的符号；

用于表征语句中的隐含知识的符号；

用于表征语句中的缺失信息的符号。

17.一种设备，其特征在于，包括：

一个或多个处理器；

存储装置，用于存储一个或多个程序，

当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求1-8中任一所述的方法。

18.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现如权利要求1-8中任一所述的方法。

用于抽取信息的方法、装置和设备

技术领域

[0001] 本申请实施例涉及计算机技术领域，具体涉及自然语言数据处理领域，尤其涉及用于抽取信息的方法、装置和设备。

背景技术

[0002] 自然语言处理，是对人们日常使用的语言进行处理，以实现与计算机之间的有效通信的技术，它是人工智能领域的一个重要研究方向。

[0003] 自然语言处理中，往往需要借助知识库来对文本语言进行理解，知识库可以提供丰富的实体知识信息，包括实体的属性信息、实体间的关系信息等。因此，如何从自然语言中提取出有效的知识信息以构造知识库，成为学术界和工业界的研究探索的热点问题。

[0004] 由于自然语言灵活多变，自动化的信息抽取具有一定的难度。尤其对于不限定领域、类别实体的开放信息抽取，需要从海量、冗余、含有噪声的不规范文本中抽取出开放类别的实体、实体关系等知识。目前学术界与工业界所采用的方案大都是基于规则的方法。其中一种基于规则的方法是对语句进行语法分析生成语法分析树，在语法分析树上施加规则来提取相关的信息。例如根据句子语法中主语、谓词和宾语的位置规则来提取对应的信息。另一种基于规则的方法是采用人工标定的规则来自动生成一些样本，利用生成的样本来训练提取器，从而实现信息提取。这些规则通常由专家来设定，针对不同类型的信息的抽取，规则的形式一般是不相同的。因此，在抽取多种类型的信息的开放信息抽取场景中，需要设定大量的规则。

发明内容

[0005] 本申请实施例提供了用于抽取信息的方法、装置和设备。

[0006] 第一方面，本申请实施例提供了一种用于抽取信息的方法，包括：获取已标注的语料库，已标注的语料库包括多条样本语句以及与各样本语句对应的已标注信息集合；分别基于已标注的语料库中的样本语句和各样本语句对应的已标注信息集合构造输入序列以及对应的输出序列；利用深度学习方法训练得出由输入序列生成输出序列的信息抽取模型；将待处理语句输入已生成的信息抽取模型，以抽取出待处理语句所包含的知识信息集合；其中，已标注信息集合包括将从对应的样本语句中抽取出的以下至少一种知识按照预设规则规范化后的信息：基于动词或介词的知识、基于名词属性的知识、实体的描述知识以及实体与概念的关系知识。

[0007] 在一些实施例中，上述利用深度学习方法训练得出由输入序列生成输出序列的信息抽取模型，包括：对输入序列中的各对象和输出序列中的各已标注信息进行向量化处理；构建信息抽取模型的神经网络，信息抽取模型的神经网络包括编码器和解码器；利用编码器将向量化处理后的输入序列映射为输入隐状态序列，利用解码器将输出序列映射为输出隐状态序列；利用解码器对输入隐状态序列进行解码，得到预测状态序列；根据预测状态序列与输出隐状态序列的差异调整神经网络的参数。

[0008] 在一些实施例中，上述利用解码器对输入隐状态序列进行解码，得到预测状态序列，包括：对于待生成的预测状态序列中的目标位置，获取解码器的隐藏层在解码得出目标位置的上一个位置的预测状态后的状态，作为解码器的当前隐状态；基于当前隐状态计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度；基于匹配度计算各输入隐状态对目标位置的预测状态的注意力权重；对各输入隐状态按照注意力权重进行加权求和得到上下文向量；基于上下文向量、输出隐状态序列中目标位置的上一个位置的输出隐状态、以及解码器的隐藏层在解码目标位置的预测状态时的状态，计算出目标位置的预测状态的概率分布；根据概率分布，确定出目标位置的预测状态。

[0009] 在一些实施例中，上述基于当前隐状态计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度，包括：对当前隐状态和输入隐状态进行第一非线性变换，得到输入隐状态与目标位置的预测状态之间的匹配度。

[0010] 在一些实施例中，上述基于当前隐状态计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度，包括：基于当前隐状态计算目标位置对应的覆盖向量，覆盖向量用于记录样本语句中已被提取的信息；对当前隐状态、输入隐状态以及目标位置的上一个位置对应的覆盖向量进行第二非线性变换，得到输入隐状态与目标位置的预测状态之间的匹配度。

[0011] 在一些实施例中，在基于当前隐状态计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度之前，上述利用解码器对输入隐状态序列进行解码，得到预测状态序列，还包括：采用自然语言处理方法提取样本语句中的语义依存关系树；对语义依存关系树中的每一条边添加一条反向边，构成路径查找图；在路径查找图中查找输入序列中任意两个对象之间的语义路径，并将该语义路径编码为语义向量；对语义向量进行变换得到语义依存关系注意力向量；上述基于当前隐状态计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度，包括：基于当前隐状态、输入隐状态以及语义依存关系注意力向量进行第三非线性变换，得到输入隐状态与目标位置的预测状态之间的匹配度。

[0012] 在一些实施例中，上述目标位置的预测状态的概率为：从对应的样本语句中拷贝一个词语作为输出序列中的一个对象的概率与从预设符号集合中选择出一个目标符号并将目标符号所表征的对象作为输出序列中的一个对象的概率之和；预设符号集合中的符号用于与样本语句中的词语联合，以完整表征样本语句中的以下知识中的一项：基于动词或介词的知识、基于名词属性的知识、实体的描述知识以及实体与概念的关系知识。

[0013] 在一些实施例中，上述预设符号集合中的符号包括：用于表征语句中分离的谓词的符号；用于表征语句中以缩减表达式表达的知识中被缩减的信息的符号；用于表征语句中的隐含知识的符号；用于表征语句中的缺失信息的符号。

[0014] 第二方面，本申请实施例提供了一种用于抽取信息的装置，包括：获取单元，用于获取已标注的语料库，已标注的语料库包括多条样本语句以及与各样本语句对应的已标注信息集合；构造单元，用于分别基于已标注的语料库中的样本语句和各样本语句对应的已标注信息集合构造输入序列以及对应的输出序列；训练单元，用于利用深度学习装置训练得出由输入序列生成输出序列的信息抽取模型；抽取单元，用于将待处理语句输入已生成

的信息抽取模型，以抽取出待处理语句所包含的知识信息集合；其中，已标注信息集合包括将从对应的样本语句中抽取出的以下至少一种知识按照预设规则规范化后的信息：基于动词或介词的知识、基于名词属性的知识、实体的描述知识以及实体与概念的关系知识。

[0015] 在一些实施例中，上述训练单元进一步用于按照如下方式训练得出由输入序列生成输出序列的信息抽取模型：对输入序列中的各对象和输出序列中的各已标注信息进行向量化处理；构建信息抽取模型的神经网络，信息抽取模型的神经网络包括编码器和解码器；利用编码器将向量化处理后的输入序列映射为输入隐状态序列，利用解码器将输出序列映射为输出隐状态序列；利用解码器对输入隐状态序列进行解码，得到预测状态序列；根据预测状态序列与输出隐状态序列的差异调整神经网络的参数。

[0016] 在一些实施例中，上述训练单元进一步用于按照如下方式得到预测状态序列：对于待生成的预测状态序列中的目标位置，获取解码器的隐藏层在解码得出目标位置的上一个位置的预测状态后的状态，作为解码器的当前隐状态；基于当前隐状态计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度；基于匹配度计算各输入隐状态对目标位置的预测状态的注意力权重；对各输入隐状态按照注意力权重进行加权求和得到上下文向量；基于上下文向量、输出隐状态序列中目标位置的上一个位置的输出隐状态、以及解码器的隐藏层在解码目标位置的预测状态时的状态，计算出目标位置的预测状态的概率分布；根据概率分布，确定出目标位置的预测状态。

[0017] 在一些实施例中，上述训练单元进一步用于按照如下方式计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度：对当前隐状态和输入隐状态进行第一非线性变换，得到输入隐状态与目标位置的预测状态之间的匹配度。

[0018] 在一些实施例中，上述训练单元进一步用于按照如下方式计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度：基于当前隐状态计算目标位置对应的覆盖向量，覆盖向量用于记录样本语句中已被提取的信息；对当前隐状态、输入隐状态以及目标位置的上一个位置对应的覆盖向量进行第二非线性变换，得到输入隐状态与目标位置的预测状态之间的匹配度。

[0019] 在一些实施例中，上述训练单元还用于在基于当前隐状态计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度之前执行如下操作：采用自然语言处理装置提取样本语句中的语义依存关系树；对语义依存关系树中的每一条边添加一条反向边，构成路径查找图；在路径查找图中查找输入序列中任意两个对象之间的语义路径，并将该语义路径编码为语义向量；对语义向量进行变换得到语义依存关系注意力向量；上述训练单元用于按照如下方式计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度：基于当前隐状态、输入隐状态以及语义依存关系注意力向量进行第三非线性变换，得到输入隐状态与目标位置的预测状态之间的匹配度。

[0020] 在一些实施例中，上述目标位置的预测状态的概率为：从对应的样本语句中拷贝一个词语作为输出序列中的一个对象的概率与从预设符号集合中选择出一个目标符号并将目标符号所表征的对象作为输出序列中的一个对象的概率之和；预设符号集合中的符号用于与样本语句中的词语联合，以完整表征样本语句中的以下知识中的一项：基于动词或

介词的知识、基于名词属性的知识、实体的描述知识以及实体与概念的关系知识。

[0021] 在一些实施例中，上述预设符号集合中的符号包括：用于表征语句中分离的谓词的符号；用于表征语句中以缩减表达式表达的知识中被缩减的信息的符号；用于表征语句中的隐含知识的符号；用于表征语句中的缺失信息的符号。

[0022] 第三方面，本申请实施例提供了一种设备，包括：一个或多个处理器；存储装置，用于存储一个或多个程序，当一个或多个程序被一个或多个处理器执行，使得一个或多个处理器实现上述用于抽取信息的方法。

[0023] 本申请实施例提供的用于抽取信息的方法、装置和设备，通过获取已标注的语料库，已标注的语料库包括多条样本语句以及与各样本语句对应的已标注信息集合；随后分别基于已标注的语料库中的样本语句和各样本语句对应的已标注信息集合构造输入序列以及对应的输出序列；之后利用深度学习方法训练得出由输入序列生成输出序列的信息抽取模型；最后将待处理语句输入已生成的信息抽取模型，以抽取出待处理语句所包含的知识信息集合；其中，已标注信息集合包括将从对应的样本语句中抽取出的以下至少一种知识按照预设规则规范化后的信息：基于动词或介词的知识、基于名词属性的知识、实体的描述知识以及实体与概念的关系知识；实现了利用一个信息抽取模型提取多种不同类型的知识信息，提升了信息抽取效率，而且基于深度学习方法训练得出的模型能够适应自然语言的表达的多样性和灵活性，避免了规则的局限性所造成的偏差，可以提升信息抽取的准确率和召回率。

附图说明

[0024] 通过阅读参照以下附图所作的对非限制性实施例详细描述，本申请的其它特征、目的和优点将会变得更明显：

[0025] 图1是本申请可以应用于其中的示例性系统架构图；

[0026] 图2是根据本申请的用于抽取信息的方法的一个实施例的流程示意图；

[0027] 图3是根据本申请的用于抽取信息的方法中利用深度学习方法训练得出信息抽取模型的一个实施例的流程示意图；

[0028] 图4是根据本申请实施例的计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度的方法的一个原理示意图；

[0029] 图5是根据本申请的用于抽取信息的装置的一个实施例的结构示意图；

[0030] 图6是适于用来实现本申请实施例的服务器或终端设备的计算机系统的结构示意图。

具体实施方式

[0031] 下面结合附图和实施例对本申请作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释相关发明，而非对该发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与有关发明相关的部分。

[0032] 需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本申请。

[0033] 图1示出了可以应用本申请的用于抽取信息的方法或装置的实施例的示例性系统

架构100。

[0034] 如图1所示，系统架构100可以包括终端设备101、102、103，网络104、以及服务器105。网络104用以在终端设备101、102、103和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

[0035] 用户110可以使用终端设备101、102、103通过网络104与服务器105交互，以接收或发送消息等。终端设备101、102、103上可以安装有各种应用，例如通讯客户端应用，语音交互类应用，文档处理类应用等，包括网页浏览器、购物类应用、搜索类应用、即时通信工具、社交平台软件等。

[0036] 终端设备101、102、103可以是具有显示屏并且支持网页浏览的各种电子设备，包括但不限于智能手机、平板电脑、电子书阅读器、智能手表、智能音箱、MP3播放器(Moving Picture Experts Group Audio Layer III，动态影像专家压缩标准音频层面3)、MP4(Moving Picture Experts Group Audio Layer IV，动态影像专家压缩标准音频层面4)播放器、膝上型便携计算机、台式计算机等等。

[0037] 服务器105可以是提供各种服务的服务器，例如对终端设备101、102、103上显示的搜索结果提供支持的搜索引擎服务器。搜索引擎服务器可以对接收到的搜索语句进行语义解析，分析搜索语句中所包含的需求信息，根据需求信息查询相关的数据，并将查询结果(例如网页链接或页面内容)反馈给终端设备。

[0038] 需要说明的是，本申请实施例所提供的用于抽取信息的方法可以由终端设备101、102、103或服务器105执行，相应地，用于抽取信息的装置可以设置于终端设备101、102、103或服务器105中。

[0039] 应该理解，图1中的终端设备、网络、服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络、服务器。

[0040] 继续参考图2，其示出了根据本申请的用于抽取信息的方法的一个实施例的流程200。该用于抽取信息的方法，包括以下步骤：

[0041] 步骤201，获取已标注的语料库。

[0042] 在这里，已标注的语料库可以包括多条样本语句以及与各样本语句对应的已标注信息集合。样本语句可以是不限定领域和实体类别的开放领域中的语句，语料库中的样本语句可以包含不同的领域、具有不同的结构化组织形式的语句。

[0043] 在本实施例中，上述用于抽取信息的方法运行于其上的电子设备(例如图1所示的终端设备或服务器)可以获取已存储的上述语料库。具体来说，上述已标注语料库可以预先构建并存储于本地，则上述电子设备可以从本地存储器中获取到该语料库；或者，上述已标注语料库可以存储于其他电子设备(例如服务器)，则上述电子设备可以通过与用于存储已标注的语料库的电子设备之间的通信连接来获取该语料库。

[0044] 已标注的语料库可以包含人工标注的语料数据，这些语料数据可以是文本形式的语句以及对这些语句中所包含的知识信息的标注结果。在本实施例中，已标注的语料库可以作为样本数据集来训练信息抽取模型，其中的语句为样本语句，对样本语句中所包含的知识信息的标注结果即为样本语句的已标注信息集合。

[0045] 已标注信息集合可以包括将从对应的样本语句中抽取出的以下至少一种知识按照预设规则规范化后的信息：基于动词或介词的知识、基于名词属性的知识、实体的描述知

识以及实体与概念的关系知识。也就是说，已标注信息集合中的信息可以是按照如下方式形成的：从样本语句中抽取出以下类型的知识：基于动词或介词的知识、基于名词属性的知识、实体的描述知识以及实体与概念的关系知识；然后按照预设规则对这些知识进行规范化表示。在这里，预设规则可以是预先定义的对语句中的各类型知识的表达方式，例如各类型知识的模板。

[0046] 上述基于动词或介词的知识可以是语句中表征动作或状态的信息的知识，基于名词属性的知识可以是语句中表征名词的属性信息的知识，实体的描述知识可以是语句中表征实体的描述信息的知识，实体与概念的关系知识可以是用于表征实体与概念之间的所属关系信息的知识。

[0047] 作为示例，语句“李白(701年－762年)，深受庄子思想影响，爽朗大方，爱饮酒作诗，喜交友，代表作有《望庐山瀑布》等著名诗歌”中包含动词“爱”、“喜”和介词“受…影响”，实体“李白”，名词“李白”，名词属性“(701年－762年)”、“代表作有《望庐山瀑布》”，描述信息“爽朗大方”，以及表征实体与概念的关系信息的“《望庐山瀑布》是著名诗歌”。

[0048] 在该示例中，上述语句包含如下类型的知识：基于动词或介词的知识：李白深受庄子思想影响，爱饮酒作诗，喜交友；基于名词属性的知识：李白的出生日期为701年，李白的逝世日期为762年，李白的代表作为《望庐山瀑布》；实体的描述知识：李白的性格爽朗大方；以及实体与概念的关系知识：《望庐山瀑布》是著名诗歌。

[0049] 上述知识中包含一些无法直接通过语义分析等方法直接从语句中提取出的知识，并且，不同语句中提取出的同一类型的知识的表达方式可能各不相同。为了表达这些无法直接提取出的知识，并统一同类型知识的表达方式，以便在后续基于深度学习的模型训练中更容易学习到信息抽取的逻辑，在标注时，可以按照预设规则对这些知识进行规范化。在这里，预设规则可以包括各类型知识的约定表达形式，还可以包括无法直接从语句中提取出的知识的约定表达形式。

[0050] 可选地，可以采用符号协助的知识表达方式来对上述知识进行规范化，这里引入的符号可以预设规则中定义的符号，这些预先定义的符号可以用于辅助语句中的词语来完整表达各类型的知识。例如引入符号“BIRTH”和“DEATH”来辅助“(701年－762年)”表达生日和卒日。

[0051] 上述预设规则可以作为标注人员对样本语句中的知识进行标注的规则。对于语句S，其中的知识可以用一个N元组来表示：(Subject，Predicate，Arg1，Arg2，…)，其中Subject表示主语，一般是实体或名词，Predicate表示谓语，一般是动词或介词，Arg1，Arg2，…表示宾语，一般可以是实体或名词。

[0052] 在一些具体的应用场景中，上述预设规则具体可以包括：尽可能用语句中原有的词语对知识进行表达；对于无法用语句中的原有词语完整地表达的知识，可以引入如下符号进行表达：

[0053] 1)引入用于表征语句中分离的谓词的符号，这里分离的谓词可以为在语句中被其他词语分离开的谓语词语或介词词语，例如上述示例中，“李白深受庄子思想影响”可以表达为(李白，深受X影响，庄子)；

[0054] 2)引入用于表征语句中以缩减表达式表达的知识中被缩减的信息的符号，缩减表达式是指不完整的表达式，例如上述示例中“李白(701年－762年)”可以表达为(李白，

BIRTH,701年)和(李白,DEATH,701年),另一个示例中“NBA(National Basketball Association)”可以表达为(NBA,=,National Basketball Association)。

[0055] 3)引入用于表征语句中的隐含知识的符号,来协助表达语句中的隐含知识,例如“英国伦敦”可以表达为(伦敦,IN,英国);

[0056] 4)用于表征语句中的缺失信息的符号来表达语句中的缺失信息,以将知识补充完整,例如上述示例中“喜交友”表达为(李白,喜,交友),“爽朗大方”表达为(李白,DESC,爽朗大方)。

[0057] 此外,在标注时,保证句子中的各知识为独立的,且所有的知识已被标注出来。并且,对于源句子中同一个谓词连接的多个名词或实体,可以表达为一个知识,例如“李白爱饮酒作诗”可以规范化表达为(李白,爱,[饮酒][作诗])。

[0058] 在对样本语句中提取出的知识进行规范化后,得到用于表征这些知识的信息,将用于表征同一样本语句中的知识的信息集合起来则可以生成对应于该样本语句的已标注信息集合。

[0059] 通过对样本语句中的知识按照上述预设规则进行规范化处理,可以表达出更丰富的知识,并保证知识的准确性。从而使得利用包含该样本语句对应的已标注信息集合训练得出的信息抽取模型可以准确地提取自然语言语句中的各类型知识,进而完善开放领域的知识库,为文本摘要、文本理解、词汇相似度等任务提供更加完备的知识信息。

[0060] 步骤202,分别基于已标注的语料库中的样本语句和各样本语句对应的已标注信息集合构造输入序列以及对应的输出序列。

[0061] 可以对上述已标注的语料库中的样本语句进行序列化处理,得到输入序列,对样本语句对应的已标注信息集合进行序列化处理,得到对应的输出序列。

[0062] 具体地,可以对样本语句进行切词,根据切词结果构造输入序列。例如样本语句“李白深受庄子思想影响,爱饮酒作诗,喜交友”的切词结果可以为“李白,深受,庄子思想,影响,爱,饮酒,作诗,喜,交友”,则输入序列S可以表示为:

[0063] $$S=\left[w_1^S,w_2^S,\cdots,w_{N_S}^S\right];\qquad(1)$$

[0064] 其中,w_i^S表示样本语句中的一个词,i=1,2,…,N_S;N_S为样本语句切词得到的词语数量。

[0065] 输出序列可以是样本语句对应的已标注信息集合中的信息构成的序列,可以表示为:

[0066] $$F=\left[w_1^F,w_2^F,\cdots,w_{N_F}^F\right];\qquad(2)$$

[0067] 其中,w_j^F表示已标注信息集合中的一个信息,j=1,2,…,N_F;N_F为已标注信息集合中的信息数量。

[0068] 步骤203,利用深度学习方法训练得出由输入序列生成输出序列的信息抽取模型。

[0069] 在本实施例中,可以构建基于深度学习的模型,然后基于上述已标注的语料库对应的输入序列和输出序列对模型进行训练。具体地,可以构建多层神经网络作为模型的架构,将输入序列输入多层神经网络进行预测,然后根据预测结果修正多层神经网络的参数。在这里,多层神经网络可以选择RNN(Recurrent Neutral Network,递归神经网络)、CNN

(Convolutional Neural Network,卷积神经网络)等。在训练过程中,上述多层神经网络可以学习由输入序列生成对应的输出序列的逻辑,也即学习到由样本语句提取其中所包含的知识的逻辑。

[0070] 步骤204,将待处理语句输入已生成的信息抽取模型,以抽取出待处理语句所包含的知识信息集合。

[0071] 在采用深度学习方法生成信息抽取模型之后,可以将待处理语句输入该信息抽取模型。信息抽取模型可以将待处理语句中所包含的知识信息抽取出来,形成知识信息集合。

[0072] 在本实施例中,待处理语句可以是一段自然语言文本,从待处理语句中抽取出的知识信息可以为与样本语句的已标注信息集合具有相同的知识表达方式。由于上述信息抽取模型的训练样本中标注了基于动词或介词的知识、基于名词属性的知识、实体的描述知识以及实体与概念的关系知识,因此该信息抽取模型可以从待处理语句中同时抽取出这四种类型的知识。

[0073] 本申请上述实施例的用于抽取信息的方法,通过获取已标注的语料库,已标注的语料库包括多条样本语句以及与各样本语句对应的已标注信息集合;随后分别基于已标注的语料库中的样本语句和各样本语句对应的已标注信息集合构造输入序列以及对应的输出序列;之后利用深度学习方法训练得出由输入序列生成输出序列的信息抽取模型;最后将待处理语句输入已生成的信息抽取模型,以抽取出待处理语句所包含的知识信息集合;其中,已标注信息集合包括将从对应的样本语句中抽取出的以下至少一种知识按照预设规则规范化后的信息:基于动词或介词的知识、基于名词属性的知识、实体的描述知识以及实体与概念的关系知识;实现了利用一个信息抽取模型提取多种不同类型的信息,提升了信息抽取效率,而且基于深度学习方法训练得出的模型能够适应自然语言的表达的多样性和灵活性,避免了规则的局限性所造成的偏差,可以提升信息抽取的准确率和召回率。

[0074] 并且,由于在训练时按照预设规则对样本语句进行了标注,形成规范化的知识信息集合,可以实现对不同领域、不同语言结构组织形式语句中的信息的抽取,并且可以有效滤除噪声的影响,从而使得该方法在开放信息抽取中能够达到良好的效果。

[0075] 在一些实施例中,可以采用序列到序列的学习的方法构造生成信息抽取模型。具体地,请参考图3,其示出了根据本申请的用于抽取信息的方法中利用深度学习方法训练得出信息抽取模型的一个实施例的流程示意图。如图3所示,上述利用深度学习方法训练得出由输入序列生成输出序列的信息抽取模型的流程300可以包括以下步骤:

[0076] 步骤301,对输入序列中的各对象和输出序列中的各已标注信息进行向量化处理。

[0077] 在本实施例中,可以构造一个词表,将输入序列S中的每个词以及输出序列F中的每个词都表示为一个独热向量(One-Hot向量),即按照词表中的词语编号或顺序,假设词表中词语数量为n,构建一个二元向量,该向量长度为n,查找输入序列S中的每个词及输出序列F中每个词在比表中的词语编号或顺序,向量中对应词语编号或顺序的元素为1,其他元素为0。

[0078] 之后,可以对输入序列S中的词语v^S的向量表示$v^S=\left(v_1^S,v_2^S,...,v_n^S\right)$和输出序列F中的词语$v^F$的向量表示$v^F=\left(v_1^F,v_2^F,...,v_n^F\right)$进行降维处理,其中$v_i^S$、$v_j^F$分别表示输入序列中的词语$v^S$的第i个元素和输出序列中的词语$v^F$的第j个元素。降维之后,可以将词语$v^S$表示为

分布式向量x_i，将词语v^F表示为分布式向量y_i。其中，$x_i = Ev_i^S$， $y_i = Ev_j^F$，E为词表矩阵。

[0079] 可选地，输入序列和输出序列中的词语的NER(Named Entity Recognition，命名实体识别)标记以及词语的词性等信息都可以通过上述相同的方式表达为分布式向量，再将每个词的所有向量拼接起来，生成该词的表示向量。

[0080] 步骤302，构建信息抽取模型的神经网络。

[0081] 其中，信息抽取模型的神经网络包括编码器和解码器。

[0082] 在本实施例中，可以构建多层神经网络作为信息抽取模型的架构。该多层神经网络可以是序列到序列的学习模型，包括编码器和解码器。

[0083] 编码器用于对输入序列进行编码，生成中间向量，解码器用于对中间向量解码，得到输入序列的预测结果。具体地，编码器可以采用RNN或DNN，包括至少一个隐层，编码器在编码时，可以结合编码器的隐层状态对当前待编码数据进行非线性变换，得到中间向量。解码器可以采用RNN或DNN，也包括至少一个隐层，解码器解码时结合已经解码的数据以及解码器的隐层的当前状态，对中间向量进行解码得到预测结果。

[0084] 步骤303，利用编码器将向量化处理后的输入序列映射为输入隐状态序列，利用解码器将输出序列映射为输出隐状态序列。

[0085] 在根据输入序列和输出序列进行迭代以调整神经网络的参数之前，可以利用信息抽取模型的神经网络来构造输入序列和输出序列的隐状态序列。在这里，隐状态序列可以是利用神经网络的隐藏层对输向量处理后的输入序列和向量化处理后的输出序列进行变换后隐藏层的状态序列。例如，输入序列中x_i的隐状态h_i^S为$h_i^S = Wx_i + b$，其中W，b为非线性变化的参数矩阵。

[0086] 在本实施例中，编码器可以包括递归神经网络，例如为2层双向递归神经网络，可以利用编码器进行非线性变换，将向量化处理后的输入序列($x_1, x_2, \cdots, x_{Ns}$)映射设为输入隐状态序列$H^S = \left(h_1^S, h_2^S, ..., h_{N_S}^S\right)$。解码器可以包括多层递归神经网络，例如为三层递归神经网络，利用解码器对向量化后的输出序列($y_1, y_2, \cdots, y_{Ns}$)映射为输出隐状态序列$H^F = \left(h_1^F, h_2^F, ..., h_{N_F}^F\right)$。

[0087] 步骤304，利用解码器对输入隐状态序列进行解码，得到预测状态序列。

[0088] 之后，可以利用解码器对输入隐状态序列$H^S = \left(h_1^S, h_2^S, ..., h_{N_S}^S\right)$进行解码。在解码时，可以根据编码器在生成输入隐状态序列中的各隐状态，生成用于辅助解码的中间语义向量，该中间语义向量可以为编码器在编码输入序列中最后一个词语时的隐状态，或者可以为编码器的各隐状态之和。

[0089] 可以依次对输入隐状态序列中的各输入隐状态，基于中间语义向量进行非线性变化，得到当前输入隐状态的预测结果。该预测结果可以是与输出隐状态序列中是输出隐状态类似的、解码器的隐藏层的预测状态。之后对该预测状态进行非线性变化输出预测结果序列。

[0090] 在对各输入隐状态进行解码之后，可以得到对应的预测状态序列。

[0091] 在本实施例中一些可选的实现方式中，上述利用解码器对输入隐状态序列进行解

码，得到预测状态序列的步骤304可以包括：

[0092] 首先，对于待生成的预测状态序列中的目标位置t，获取解码器的隐藏层在解码得出目标位置t的上一个位置t-1的预测状态后的状态s_{t-1}，作为解码器的当前隐状态。

[0093] 然后，基于当前隐状态s_{t-1}计算输入隐状态序列中各输入隐状态h_j^S与待生成的预测状态序列$H^O=\left(h_1^O,h_2^O,...,h_{N_O}^O\right)$中目标位置t的预测状态$h_t^O$之间的匹配度$e_{tj}$，在这里，匹配度计算方法也可以称为对齐模型，匹配度e_{tj}可以表示目标位置t与输入序列的位置j的对齐程度。

[0094] 在一些可选的实现方式中，可以按照如下方式计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度：对当前隐状态和输入隐状态进行第一非线性变换，得到输入隐状态与目标位置的预测状态之间的匹配度。具体可以按照式(3)计算上述匹配度：

[0095] $$e_{tj}=a\left(s_{t-1},h_j^S\right)=v_a^T\tanh\left(W_a s_{t-1}+U_a h_j^S\right) \quad (3)$$

[0096] 其中，v_a^T是解码器的参数向量，W_a和U_a是解码器的参数矩阵，tanh为第一非线性变换的变换函数。

[0097] 在另一些可选的实现方式中，可以引入覆盖机制(Coverage机制)，来改善“欠提取”和“过提取”的问题，可以按照如下方式计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度：基于当前隐状态计算目标位置对应的覆盖向量，覆盖向量用于记录样本语句中已被提取的信息；对当前隐状态、输入隐状态以及目标位置的上一个位置对应的覆盖向量进行第二非线性变换，得到输入隐状态与目标位置的预测状态之间的匹配度。

[0098] 在上述引入覆盖机制的实现方式中，可以对每个词汇(位置t的词语)设置一个覆盖向量m_j^t，该覆盖向量m_j^t可以按照如下方法更新：

[0099] $$m_j^t=\mu\left(m_j^{t-1},\alpha_{tj},h_j^S,s_{t-1}\right)=\left(1-z_i\right)\circ m_j^{t-1}+z_j\circ\tilde{m}_j^t \quad (4)$$

[0100] 其中，

[0101] $$\tilde{m}_j^t=\tanh\left(W_h h_j^S+u_\alpha\alpha_{tj}+W_s s_{t-1}+U_m\left[r_i\circ m_j^{t-1}\right]\right) \quad (5)$$

[0102] 其中，“∘”为矩阵按元素相乘操作，z_i为更新门，r_i为重置门，分别按照式(6)、(7)计算：

[0103] $$z_i=\sigma\left(W_h^z h_j^S+u_\alpha^z\alpha_{tj}+W_s^z s_{t-1}+U_m^z m_j^{t-1}\right) \quad (6)$$

[0104] $$r_i=\sigma\left(W_h^r h_j^S+u_\alpha^r\alpha_{tj}+W_s^r s_{t-1}+U_m^r m_j^{t-1}\right) \quad (7)$$

[0105] 其中，W_h^z、W_s^z、U_m^z、W_h^r、W_s^r、U_m^r、W_h、W_s和U_m均为神经网络的参数矩阵，u_α、u_α^z、u_α^r均为神经网络的参数向量，σ表示激活函数，可以是非线性变换函数。

[0106] 在这里，覆盖向量可以记录在信息抽取过程中源句子(训练时为样本语句)中的词语被拷贝以形成知识信息的历史记录，从而可以帮助解码器进行当前的决策。这时，上述匹

配度可以按照式(8)计算：

[0107] $$e_{tj}=a\left(s_{t-1},h_j^S,m_j^{t-1}\right)=v_a^T\tanh\left(W_a s_{t-1}+U_a h_j^S+V_a m_j^{t-1}\right) \quad (8)$$

[0108] 其中，V_a为神经网络中的参数矩阵，tanh为第二非线性变换的变换函数。

[0109] 从式(8)可以看出，匹配度e_{tj}是由当前隐状态s_{t-1}、输入隐状态h_j^S以及目标位置的上一个位置对应的覆盖向量m_j^{t-1}进行第二非线性变换后得到的，这样，可以结合已解码出的知识信息对当前待解码知识进行提取，可以避免知识信息的重复提取，同时可以保证不会遗漏知识信息。

[0110] 在另一些可选的实现方式中，在基于当前隐状态计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度之前，上述利用解码器对输入隐状态序列进行解码，得到预测状态序列的步骤，还可以包括语义依存关系分析步骤，语义依存关系分析步骤可以包括：

[0111] 首先采用自然语言处理方法提取样本语句中的语义依存关系树，具体可以利用语义分析方法分析句子各个语言单位之间的语义关联，并将语义关联以依存的树状结构呈现。其中，语义依存关系可以包括主要语义角色、事件关系以及语义依附标记，关系类型可以包括主谓关系、动宾关系、间宾关系、前置宾语、动补结构、介宾关系等。语义关系树中的每一条边具有一个标签，表示一种关系。

[0112] 然后，可以对语义依存关系树中的每一条边添加一条反向边，构成路径查找图。具体地可以在原边的标签前添加一个负号形成反向边的标签。

[0113] 而后，在路径查找图中查找输入序列中任意两个对象之间的语义路径，并将该语义路径编码为语义向量。具体可以在路径查找图中找出样本语句中两个词汇$(w_{i'}^S,\ w_{j'}^S)$之间的最短路径作为语义路径，可以采用RNN等方法将该语义路径编码成为语义向量n_{ij}。

[0114] 之后，对语义向量进行变换得到语义依存关系注意力向量。具体地，语义依存关系注意力向量$u_j^t=g\left(\tilde{u}_j^t\right)$，其中，g是诸如GRU(Gated Recurrent Unit，门限递归单元)的门限操作，可以是由当前隐状态和输入隐状态定义的一个门函数，$\tilde{u}_j^t$可以为：

[0115] $$\tilde{u}_j^t=\sum_i p^*\left(w_t^F=w_i^S\right)n_{ij} \quad (9)$$

[0116] 其中，p*是目标位置的预测状态h_t^O的概率分布p的锐化结果。在这里，可以对p进行k次方运算后归一化得到p*，使得接近于1的概率值更接近1，而较小的概率则趋于0。

[0117] 这时，上述基于当前隐状态计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度的步骤，可以包括：基于当前隐状态、输入隐状态以及语义依存关系注意力向量进行第三非线性变换，得到输入隐状态与目标位置的预测状态之间的匹配度。具体来说，匹配度可以按照式(10)计算：

[0118] $$e_{tj}=a\left(s_{t-1},h_j^S,m_j^{t-1}\right)=v_a^T\tanh\left(W_a s_{t-1}+U_a h_j^S+V_a m_j^{t-1}+D_a u_j^{t-1}\right) \quad (10)$$

[0119] 其中，D_a也是神经网络中的参数矩阵，用于对语义依存关系注意力向量进行变换，tanh为第三非线性变换的变换函数。

[0120] 通过引入语义依存关系注意力向量，神经网络可以自动根据源句子的语义确定在解码时所依赖的已解码词汇和待解码词汇，在训练过程中有效地融入各样本语句的语义依存关系，进一步提升信息抽取模型的准确性。

[0121] 图4示出了根据本申请实施例的计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度的方法的一个原理示意图。如图4所示，信息抽取模型的神经网络通过对输入隐状态和当前隐状态进行诸如tanh函数的非线性变换，可以得出匹配度；在另一中场景中，可以通过对输入隐状态、当前隐状态以及覆盖向量进行诸如tanh函数的非线性变换，得到匹配度；还可以通过对输入隐状态、当前隐状态以及语义依存关系注意力向量进行诸如tanh函数的非线性变换，得到匹配度；或者通过对输入隐状态、当前隐状态、覆盖向量以及语义依存关系注意力向量进行诸如tanh函数的非线性变换，得到匹配度。

[0122] 可以基于匹配度e_{tj}计算各输入隐状态对目标位置t的预测状态h_t^O的注意力权重，具体地，输入隐状态序列中第j个位置的输入隐状态h_j^S对目标位置t的预测状态h_t^O的注意力权重α_{tj}为：

[0123] $$\alpha_{tj}=\frac{\exp\left(e_{tj}\right)}{\sum_{k=1}^{N_S}\exp\left(e_{tk}\right)} \qquad (11)$$

[0124] 注意力权重α_{tj}也可以称为注意力回应强度，可以是前面的预测状态或输出状态对当前输入隐状态解码的影响力强度。

[0125] 接着，可以对各输入隐状态按照注意力权重进行加权求和得到上下文向量，具体地，上下文向量c_t为：

[0126] $$c_t=\sum_{j=1}^{N_S}\alpha_{tj}h_j^S \qquad (12)$$

[0127] 在这里，上下文向量c_t是一个动态向量，不同目标位置的上下文向量不相同。

[0128] 之后，使用上下文向量c_t、输出隐状态序列中目标位置的上一个位置t-1的输出隐状态h_{t-1}^F以及解码器的隐藏层在解码目标位置的预测状态时的状态s_t，计算目标位置的预测状态h_t^O的概率分布。

[0129] 具体地，目标位置的预测状态h_t^O的概率分布可以按照式(13)计算得出：

[0130] $$p\left(h_t^O\middle|\left\{h_1^O,h_2^O,\cdots,h_{t-1}^O\right\},c_t\right)=f\left(h_{t-1}^F,s_t,c_t\right) \qquad (13)$$

[0131] 其中，f是一个词汇生成模型，可以是一个非线性变换函数。

[0132] 最后，根据概率分布，确定出目标位置的预测状态。可以在预设的词汇表中查找使得式(13)的概率p值最大的词汇作为目标位置t的预测状态。在这里，预设的词汇表可以为预先定义的、用于表示上各种知识的词汇。在查找使得概率p值最大的词汇时，可以运用贪心算法或集束搜索(Beam Search，一种启发式搜索算法)得到最优的选择。

[0133] 从式(3)-(13)可以看出，预测状态与输入隐状态序列中的各隐状态h_j^S、根据解码器的隐藏层在解码得出目标位置的上一个位置的预测状态后的状态确定的注意力权重α_{tj}

以及输出隐状态序列中目标位置的上一个位置的隐状态h_{t-1}^{F}相关，则本实施例的信息抽取模型的训练方法可以有效地结合上下文信息，并且考虑不同位置的上下文对当前位置的解码的影响力大小，能够提升模型的准确性。

[0134] 在一些可选的实现方式中，可以采用受约束的copy(拷贝)机制对上述概率分布p的计算方法进行优化。该受约束的copy机制用于约束解码过程中遵循“尽量使用源句子中的词汇来表达知识”的规则，以避免解码器生成的知识信息与源句子语义差异过大。

[0135] 具体地，可以首先定义上述按照预设规则表达的知识中引入的用于辅助表达知识的预设符号集合，也即，该预设符号集合中的符号用于与样本语句中的词语联合，以完整表征样本语句中的以下知识中的一项：基于动词或介词的知识、基于名词属性的知识、实体的描述知识以及实体与概念的关系知识。

[0136] 可选地，上述预设符号集合中的符号可以包括但不限于：用于表征语句中分离的谓词的符号；用于表征语句中以缩减表达式表达的知识中被缩减的信息的符号；用于表征语句中的隐含知识的符号；以及用于表征语句中的缺失信息的符号。

[0137] 这时，上述目标位置的预测状态的概率可以为：从对应的样本语句中拷贝一个词语作为输出序列中的一个对象的概率与从预设符号集合中选择出一个目标符号并将目标符号所表征的对象作为输出序列中的一个对象的概率之和。也就是说，

[0138] $$p\left(h_t^O\middle|\left\{h_1^O,h_2^O,\cdots,h_{t-1}^O\right\},c_t\right)=p_X\left(h_t^O\middle|\left\{h_1^O,h_2^O,\cdots,h_{t-1}^O\right\},c_t\right)+p_K\left(h_t^O\middle|\left\{h_1^O,h_2^O,\cdots,h_{t-1}^O\right\},c_t\right)\qquad(14)$$

[0139] 其中，p_X为从从对应的样本语句中拷贝一个词语作为输出序列中的一个对象的概率，p_K为从预设符号集合中选择出一个目标符号并将目标符号所表征的对象作为输出序列中的一个对象的概率，一种可选的计算方式如下：

[0140] $$p_X\left(h_t^O\middle|\left\{h_1^O,h_2^O,\cdots,h_{t-1}^O\right\},c_t\right)=\frac{1}{z}\exp\left(\sigma\left(\left(h_j^S\right)^T W_c\right)s_t\right)\qquad(15)$$

[0141] $$p_K\left(h_t^O\middle|\left\{h_1^O,h_2^O,\cdots,h_{t-1}^O\right\},c_t\right)=\frac{1}{z}\exp\left(v_i^T W_O s_t\right)\qquad(15)$$

[0142] 其中，W_c、W_0是参数矩阵，v_i是参数向量。

[0143] 步骤305，根据预测状态序列与输出隐状态序列的差异调整神经网络的参数。

[0144] 在生成各目标位置的预测状态，进而得到预测状态序列之后，可以比对预测状态序列与输出隐状态序列，根据二者之间的差异对信息抽取模型的神经网络中的各参数进行调整。

[0145] 具体地，上述预测状态序列与输出隐状态序列之间的差异可以由损失函数来表示，然后计算损失函数关于信息抽取模型的神经网络中各参数的梯度，采用梯度下降法来更新各参数。

[0146] 进一步地，如图3所示(虚线箭头所示流程)，可以将输入隐状态序列重新输入更新参数后的神经网络，得到新的预测状态序列，之后在重复上述比对预测状态序列和输出状态序列和更新参数的步骤，直到损失函数符合预设的收敛条件。

[0147] 基于如图3所示的流程，本实施例的利用深度学习方法训练得出由输入序列生成输出序列的信息抽取模型的方法可以精准地提取上下文信息，并且输入隐状态序列、输出隐状态序列的长度可变，能够灵活地适应不同的语句。由此，采用基于图3所示方法生成的

信息抽取模型可以精准地提取出待处理语句中的各类型知识。实验结果中，采用图3所示流程生成的信息抽取模型在准确率、召回率和F值方面均优于现有的基于规则的方法。

[0148] 进一步参考图5，作为对上述图2所示方法的实现，本申请提供了一种用于抽取信息的装置的一个实施例，该装置实施例与图2所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

[0149] 如图5所示，本实施例的用于抽取信息的装置500可以包括：获取单元501、构造单元502、训练单元503以及抽取单元504。其中，获取单元501可以用于获取已标注的语料库，已标注的语料库包括多条样本语句以及与各样本语句对应的已标注信息集合；构造单元502可以用于分别基于已标注的语料库中的样本语句和各样本语句对应的已标注信息集合构造输入序列以及对应的输出序列；训练单元503可以用于利用深度学习装置训练得出由输入序列生成输出序列的信息抽取模型；抽取单元504可以用于将待处理语句输入已生成的信息抽取模型，以抽取出待处理语句所包含的知识信息集合；其中，已标注信息集合包括将从对应的样本语句中抽取出的以下至少一种知识按照预设规则规范化后的信息：基于动词或介词的知识、基于名词属性的知识、实体的描述知识以及实体与概念的关系知识。

[0150] 在本实施例中，已标注语料库可以预先构建并存储于本地，则获取单元501可以从本地存储器中获取到该语料库；或者，上述已标注语料库可以存储于其他电子设备(例如服务器)，则获取单元501可以通过与用于存储已标注的语料库的电子设备之间的通信连接来获取该语料库。在这里，已标注的语料库可以包含人工标注的语料数据，这些语料数据可以是文本形式的语句以及对这些语句中所包含的知识信息的标注结果。并且，这些知识信息可以是按照预设规则对语句中包含的知识规范化后的信息，预设规则可以作为标注人员对样本语句中的知识行标注的规则。

[0151] 构造单元502可以对样本语句进行切词，基于切词得到的词语构造输入序列，还可以将样本语句对应的已标注信息集合构造为输出序列。

[0152] 训练单元503可以构建基于深度学习的模型，然后基于上述已标注的语料库对应的输入序列和输出序列对模型进行训练。例如可以构建多层神经网络，将输入序列作为多层神经网络的输入数据，多层神经网络可以根据输入数据进行预测，得到预测结果，将输出序列作为标注结果，计算预测结果与标注结果之间的差异，然后调整多层神经网络的参数使差异缩小到一定的范围，从而得出已训练好的信息抽取模型。

[0153] 抽取单元504可以利用训练单元503得出的信息抽取模型对待处理语句进行信息抽取，具体可以将待处理语句输入信息抽取模型，得到待处理中的知识信息集合。

[0154] 在一些实施例中，上述训练单元503可以进一步用于按照如下方式训练得出由输入序列生成输出序列的信息抽取模型：对输入序列中的各对象和输出序列中的各已标注信息进行向量化处理；构建信息抽取模型的神经网络，信息抽取模型的神经网络包括编码器和解码器；利用编码器将向量化处理后的输入序列映射为输入隐状态序列，利用解码器将输出序列映射为输出隐状态序列；利用解码器对输入隐状态序列进行解码，得到预测状态序列；根据预测状态序列与输出隐状态序列的差异调整神经网络的参数。

[0155] 在进一步的实施例中，训练单元503可以进一步用于按照如下方式得到预测状态序列：对于待生成的预测状态序列中的目标位置，获取解码器的隐藏层在解码得出目标位置的上一个位置的预测状态后的状态，作为解码器的当前隐状态；基于当前隐状态计算输

入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度;基于匹配度计算各输入隐状态对目标位置的预测状态的注意力权重;对各输入隐状态按照注意力权重进行加权求和得到上下文向量;基于上下文向量、输出隐状态序列中目标位置的上一个位置的输出隐状态、以及解码器的隐藏层在解码目标位置的预测状态时的状态,计算出目标位置的预测状态的概率分布;根据概率分布,确定出目标位置的预测状态。

[0156] 在进一步的实施例中,训练单元503可以进一步用于按照如下方式计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度:对当前隐状态和输入隐状态进行第一非线性变换,得到输入隐状态与目标位置的预测状态之间的匹配度。

[0157] 在进一步的实施例中,训练单元503可以进一步用于按照如下方式计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度:基于当前隐状态计算目标位置对应的覆盖向量,覆盖向量用于记录样本语句中已被提取的信息;对当前隐状态、输入隐状态以及目标位置的上一个位置对应的覆盖向量进行第二非线性变换,得到输入隐状态与目标位置的预测状态之间的匹配度。

[0158] 在进一步的实施例中,训练单元503还可以用于在基于当前隐状态计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度之前执行如下操作:采用自然语言处理装置提取样本语句中的语义依存关系树;对语义依存关系树中的每一条边添加一条反向边,构成路径查找图;在路径查找图中查找输入序列中任意两个对象之间的语义路径,并将该语义路径编码为语义向量;对语义向量进行变换得到语义依存关系注意力向量。并且,训练单元503可以用于按照如下方式计算输入隐状态序列中各输入隐状态与待生成的预测状态序列中目标位置的预测状态之间的匹配度:基于当前隐状态、输入隐状态以及语义依存关系注意力向量进行第三非线性变换,得到输入隐状态与目标位置的预测状态之间的匹配度。

[0159] 在一些实施例中,上述目标位置的预测状态的概率可以为:从对应的样本语句中拷贝一个词语作为输出序列中的一个对象的概率与从预设符号集合中选择出一个目标符号并将目标符号所表征的对象作为输出序列中的一个对象的概率之和;预设符号集合中的符号用于与样本语句中的词语联合,以完整表征样本语句中的以下知识中的一项:基于动词或介词的知识、基于名词属性的知识、实体的描述知识以及实体与概念的关系知识。

[0160] 在进一步的实施例中,上述预设符号集合中的符号可以包括:用于表征语句中分离的谓词的符号;用于表征语句中以缩减表达式表达的知识中被缩减的信息的符号;用于表征语句中的隐含知识的符号;用于表征语句中的缺失信息的符号。

[0161] 本申请实施例的用于抽取信息的装置500,获取单元获取已标注的语料库,已标注的语料库包括多条样本语句以及与各样本语句对应的已标注信息集合;构造单元分别基于已标注的语料库中的样本语句和各样本语句对应的已标注信息集合构造输入序列以及对应的输出序列;训练单元利用深度学习方法训练得出由输入序列生成输出序列的信息抽取模型;抽取单元将待处理语句输入已生成的信息抽取模型,以抽取出待处理语句所包含的知识信息集合;其中,已标注信息集合包括将从对应的样本语句中抽取出的以下至少一种知识按照预设规则规范化后的信息:基于动词或介词的知识、基于名词属性的知识、实体的

描述知识以及实体与概念的关系知识;实现了利用一个信息抽取模型提取多种不同类型的知识信息,提升了信息抽取效率,而且基于深度学习方法训练得出的模型能够适应自然语言的表达的多样性和灵活性,避免了规则的局限性所造成的偏差,可以提升信息抽取的准确率和召回率。

[0162] 应当理解,装置500中记载的诸单元可以与参考图2和图3描述的用于抽取信息的方法中的各个步骤相对应。由此,上文针对用于抽取信息的方法描述的操作和特征同样适用于装置500及其中包含的单元,在此不再赘述。

[0163] 下面参考图6,其示出了适于用来实现本申请实施例的终端设备或服务器的计算机系统600的结构示意图。图6示出的终端设备或服务器仅仅是一个示例,不应对本申请实施例的功能和使用范围带来任何限制。

[0164] 如图6所示,计算机系统600包括中央处理单元(CPU)601,其可以根据存储在只读存储器(ROM)602中的程序或者从存储部分608加载到随机访问存储器(RAM)603中的程序而执行各种适当的动作和处理。在RAM 603中,还存储有系统600操作所需的各种程序和数据。CPU 601、ROM 602以及RAM 603通过总线604彼此相连。输入/输出(I/O)接口605也连接至总线604。

[0165] 以下部件连接至I/O接口605:包括键盘、鼠标等的输入部分606;包括诸如阴极射线管(CRT)、液晶显示器(LCD)等以及扬声器等的输出部分607;包括硬盘等的存储部分608;以及包括诸如LAN卡、调制解调器等的网络接口卡的通信部分609。通信部分609经由诸如因特网的网络执行通信处理。驱动器610也根据需要连接至I/O接口605。可拆卸介质611,诸如磁盘、光盘、磁光盘、半导体存储器等等,根据需要安装在驱动器610上,以便于从其上读出的计算机程序根据需要被安装入存储部分608。

[0166] 特别地,根据本公开的实施例,上文参考流程图描述的过程可以被实现为计算机软件程序。例如,本公开的实施例包括一种计算机程序产品,其包括承载在计算机可读介质上的计算机程序,该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中,该计算机程序可以通过通信部分609从网络上被下载和安装,和/或从可拆卸介质611被安装。在该计算机程序被中央处理单元(CPU)601执行时,执行本申请的方法中限定的上述功能。需要说明的是,本申请所述的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于:具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本申请中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本申请中,计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含

的程序代码可以用任何适当的介质传输，包括但不限于：无线、电线、光缆、RF等等，或者上述的任意合适的组合。

[0167] 附图中的流程图和框图，图示了按照本申请各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，该模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0168] 描述于本申请实施例中所涉及到的单元可以通过软件的方式实现，也可以通过硬件的方式来实现。所描述的单元也可以设置在处理器中，例如，可以描述为：一种处理器包括获取单元、构造单元、训练单元和抽取单元。其中，这些单元的名称在某种情况下并不构成对该单元本身的限定，例如，获取单元还可以被描述为“获取已标注的语料库的单元”。

[0169] 作为另一方面，本申请还提供了一种计算机可读介质，该计算机可读介质可以是上述实施例中描述的装置中所包含的；也可以是单独存在，而未装配入该装置中。上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被该装置执行时，使得该装置：获取已标注的语料库，所述已标注的语料库包括多条样本语句以及与各样本语句对应的已标注信息集合；分别基于所述已标注的语料库中的样本语句和各所述样本语句对应的已标注信息集合构造输入序列以及对应的输出序列；利用深度学习方法训练得出由所述输入序列生成所述输出序列的信息抽取模型；将待处理语句输入已生成的信息抽取模型，以抽取出所述待处理语句所包含的知识信息集合；其中，所述已标注信息集合包括将从对应的样本语句中抽取出的以下至少一种知识按照预设规则规范化后的信息：基于动词或介词的知识、基于名词属性的知识、实体的描述知识以及实体与概念的关系知识。

[0170] 以上描述仅为本申请的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本申请中所涉及的发明范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离上述发明构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本申请中公开的(但不限于)具有类似功能的技术特征进行互相替换而形成的技术方案。

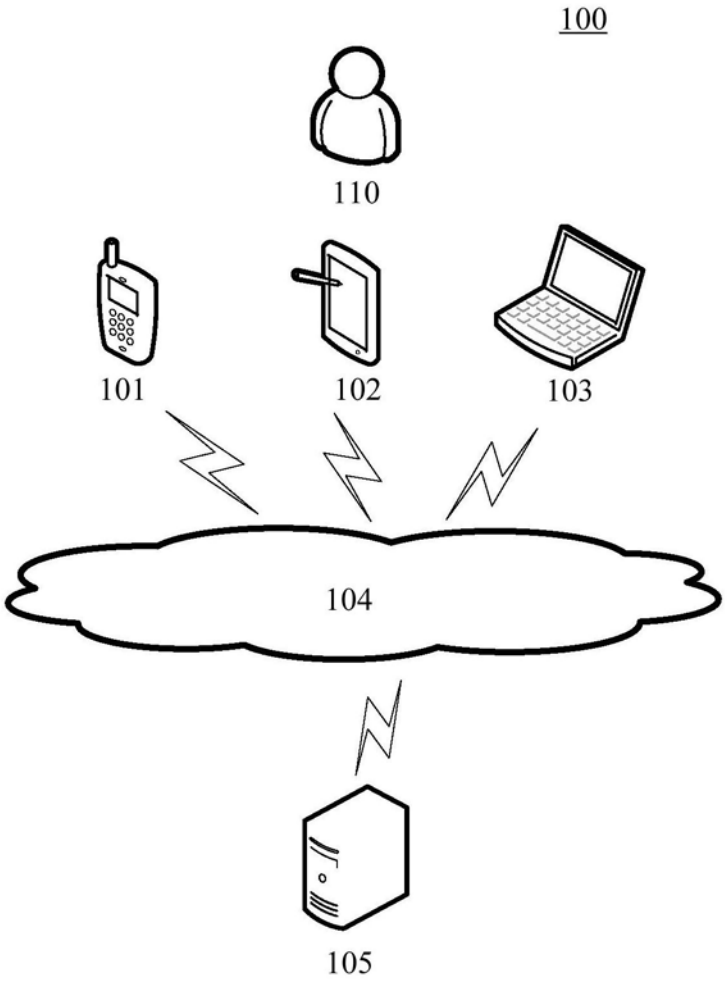


图1

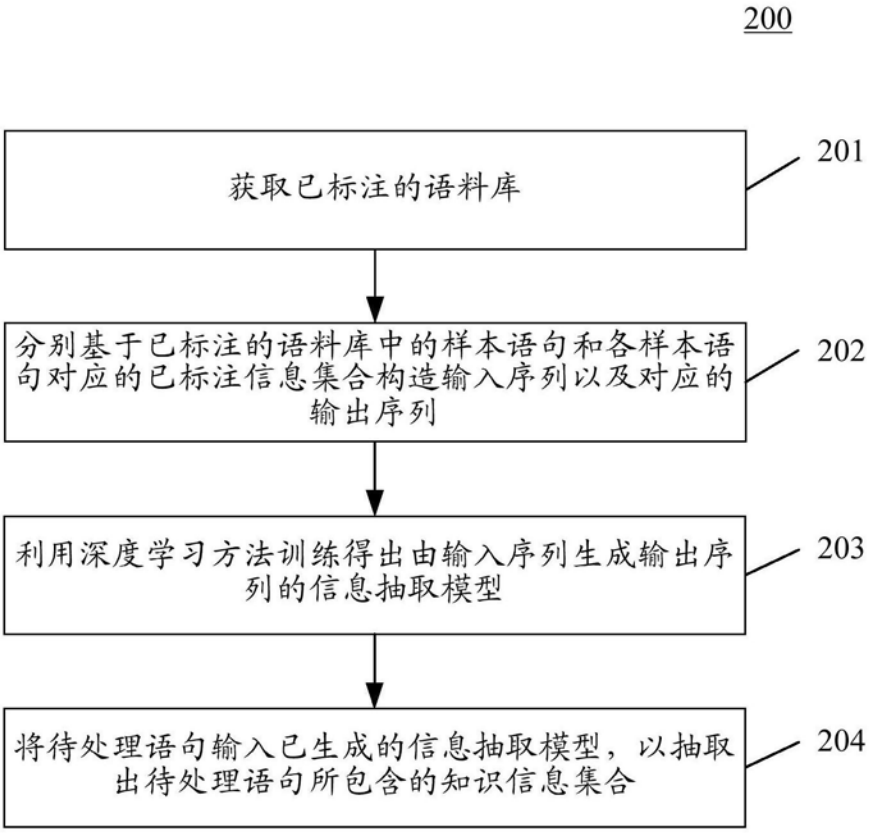

200
获取已标注的语料库
201
分别基于已标注的语料库中的样本语句和各样本语句对应的已标注信息集合构造输入序列以及对应的输出序列
202
利用深度学习方法训练得出由输入序列生成输出序列的信息抽取模型
203
将待处理语句输入已生成的信息抽取模型，以抽取出待处理语句所包含的知识信息集合
204

图2

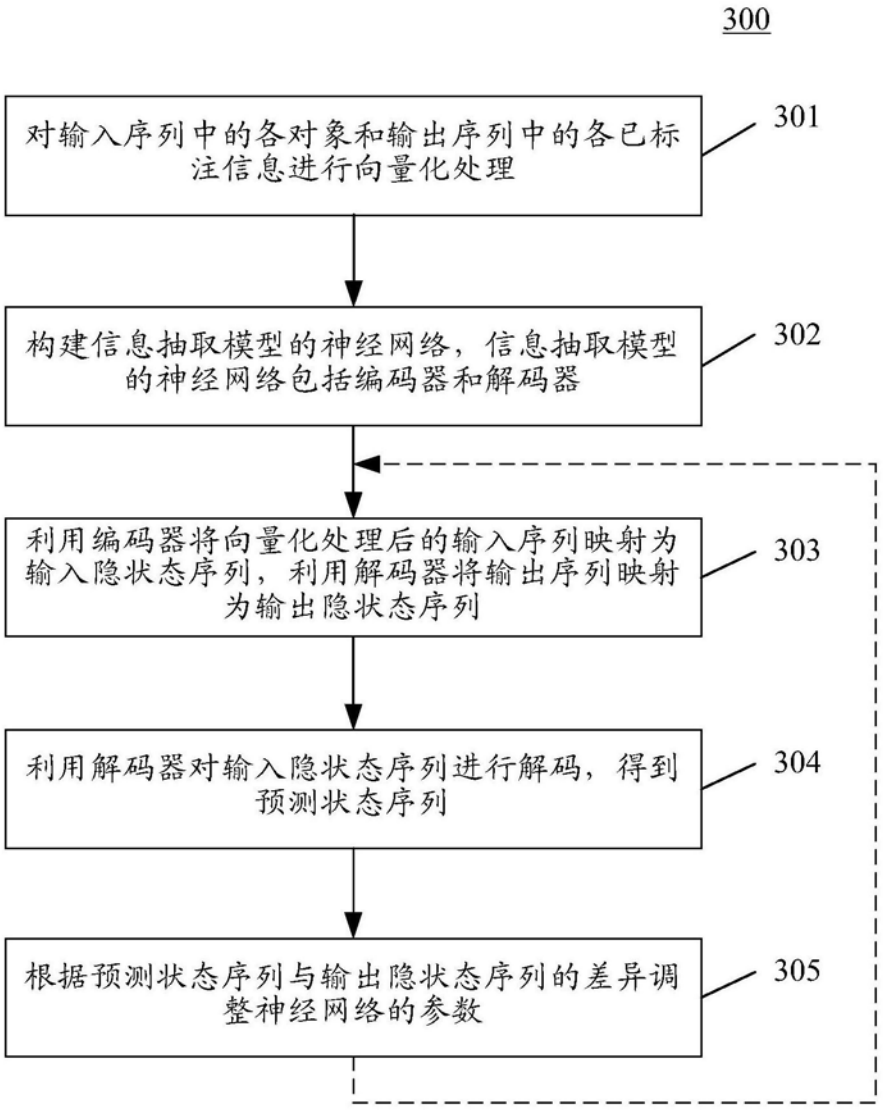

300
对输入序列中的各对象和输出序列中的各已标注信息进行向量化处理
301
构建信息抽取模型的神经网络，信息抽取模型的神经网络包括编码器和解码器
302
利用编码器将向量化处理后的输入序列映射为输入隐状态序列，利用解码器将输出序列映射为输出隐状态序列
303
利用解码器对输入隐状态序列进行解码，得到预测状态序列
304
根据预测状态序列与输出隐状态序列的差异调整神经网络的参数
305

图3

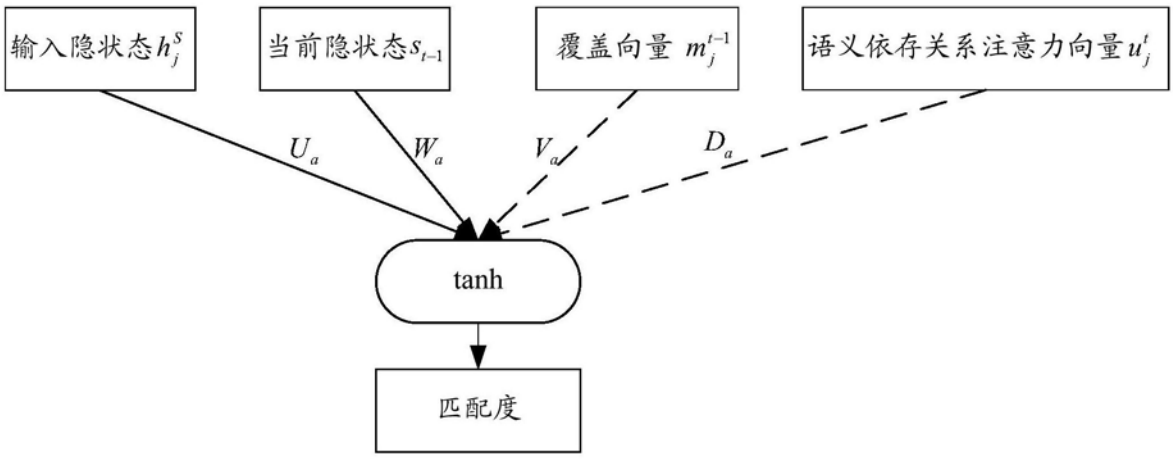

输入隐状态 h_j^S
当前隐状态 s_{t-1}
覆盖向量 m_j^{t-1}
语义依存关系注意力向量 u_j^t
U_a
W_a
V_a
D_a
tanh
匹配度

图4

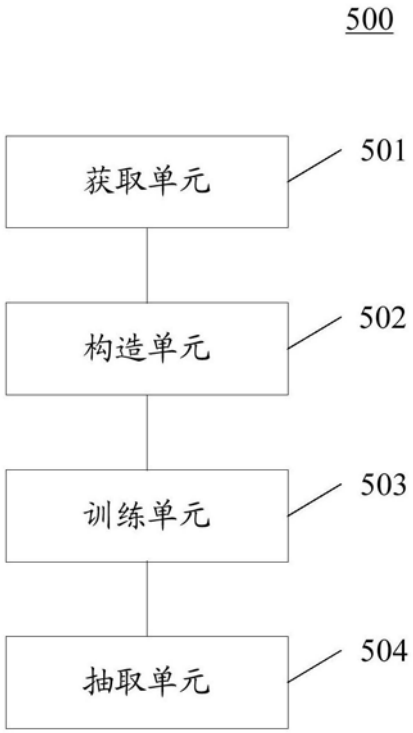

500
获取单元
501
构造单元
502
训练单元
503
抽取单元
504

图5

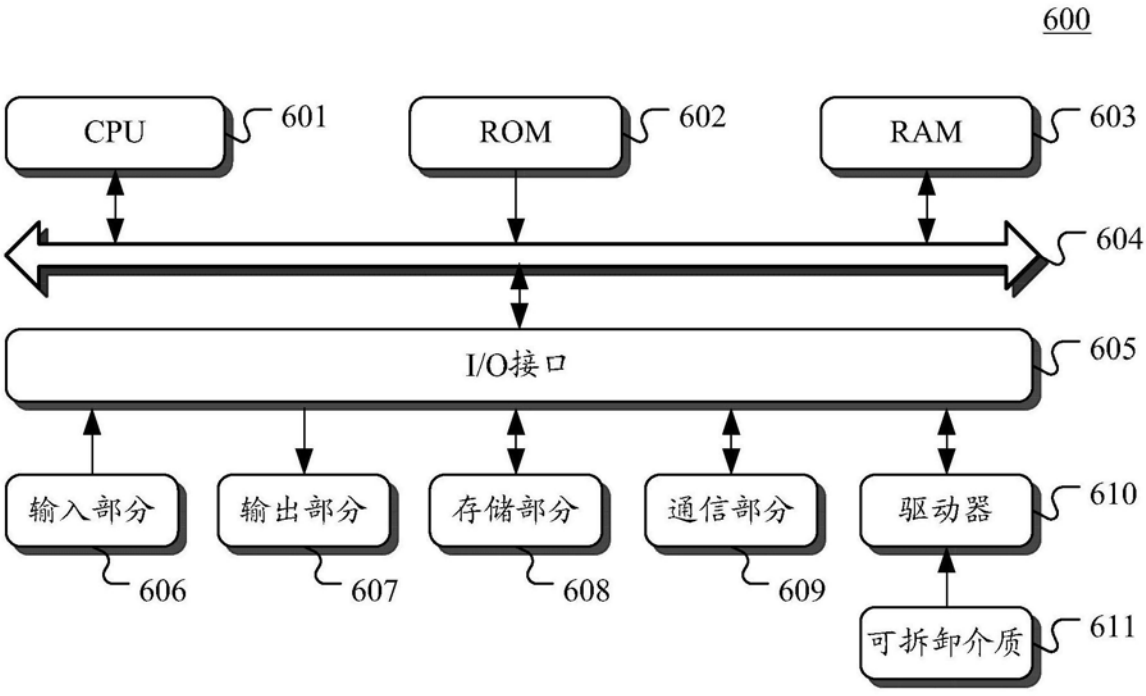

600
CPU
601
ROM
602
RAM
603
604
I/O接口
605
输入部分
606
输出部分
607
存储部分
608
通信部分
609
驱动器
610
可拆卸介质
611

图6