(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111353310 B
(45)授权公告日 2023.08.11

(21)申请号 202010127101.0

(22)申请日 2020.02.28

(65)同一申请的已公布的文献号
申请公布号 CN 111353310 A

(43)申请公布日 2020.06.30

(73)专利权人 腾讯科技(深圳)有限公司
地址 518000 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 慕福楠 吴晨光 王莉峰

(74)专利代理机构 北京派特恩知识产权代理有限公司 11270
专利代理师 王姗姗 张颖玲

(51)Int.Cl.
G06F 40/295(2020.01)
G06F 40/289(2020.01)
G06F 40/30(2020.01)
G06F 16/35(2019.01)
G06F 16/36(2019.01)
G06N 3/0464(2023.01)
G06N 3/0455(2023.01)
G06N 3/047(2023.01)
G06N 3/048(2023.01)

(56)对比文件
CN 106547733 A,2017.03.29
CN 110569366 A,2019.12.13
CN 109388793 A,2019.02.26
CN 110209812 A,2019.09.06
CN 110502738 A,2019.11.26
US 2019179897 A1,2019.06.13
US 2015286629 A1,2015.10.08

审查员 武晓冬

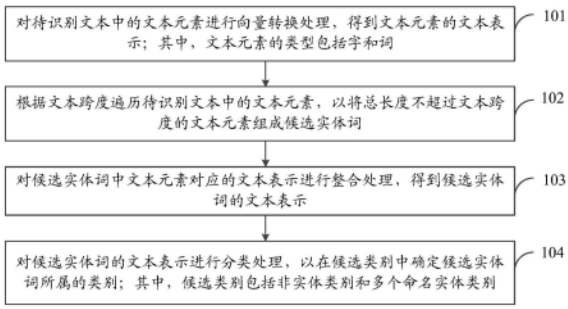
权利要求书3页 说明书18页 附图9页

(54)发明名称

基于人工智能的命名实体识别方法、装置及电子设备

(57)摘要

本发明提供了一种基于人工智能的命名实体识别方法、装置、电子设备及计算机可读存储介质;方法包括:对待识别文本中的文本元素进行向量转换处理,得到所述文本元素的文本表示;其中,所述文本元素的类型包括字和词;根据文本跨度遍历所述待识别文本中的文本元素,以将总长度不超过所述文本跨度的文本元素组成候选实体词;对所述候选实体词中文本元素对应的文本表示进行整合处理,得到所述候选实体词的文本表示;对所述候选实体词的文本表示进行分类处理,以在候选类别中确定所述候选实体词所属的类别;其中,所述候选类别包括非实体类别和多个命名实体类别。通过本发明,能够提升命名实体识别的效率和灵活性。



CN 111353310 B

1.一种基于人工智能的命名实体识别方法,其特征在于,包括:

对待识别文本中的文本元素进行向量转换处理,得到所述文本元素的文本表示;其中,所述文本元素的类型包括字和词;

根据文本跨度遍历所述待识别文本中的文本元素,以将总长度不超过所述文本跨度的文本元素组成候选实体词;

对所述候选实体词中文本元素对应的文本表示进行整合处理,得到所述候选实体词的文本表示;

对所述候选实体词的文本表示进行分类处理,以在候选类别中确定所述候选实体词所属的类别;其中,所述候选类别包括非实体类别和多个命名实体类别;

其中,所述根据文本跨度遍历所述待识别文本中的文本元素,以将总长度不超过所述文本跨度的文本元素组成候选实体词,包括:

执行以下任意一种处理:

对所述待识别文本中的文本元素进行第一层次遍历循环,其中,所述第一层次遍历循环包括多次第一层次遍历;将每个所述第一层次遍历得到的文本元素确定为起点元素;针对在每个所述第一层次遍历中确定的起点元素,执行包括多次第二层次遍历的第二层次遍历循环:根据在所述第二层次遍历循环中已经执行的第二层次遍历的次数,确定同步增大或缩小的扫描范围,根据所述扫描范围对所述待识别文本中的文本元素进行从所述起点元素开始的第二层次遍历,并将所述起点元素和所述第二层次遍历得到的文本元素组合为候选实体词,直至得到的候选实体词的长度等于所述文本跨度;或

根据多个不同长度的卷积窗口执行以下操作:在所述待识别文本中执行所述卷积窗口的滑动操作,且每次滑动操作的幅度为一个文本元素;将每次滑动后所述卷积窗口所覆盖的文本元素组合为候选实体词;其中,所述卷积窗口的长度小于或等于所述文本跨度。

2.根据权利要求1所述的命名实体识别方法,其特征在于,在所述将所述起点元素和所述第二层次遍历得到的文本元素组合为候选实体词,直至得到的候选实体词的长度等于所述文本跨度之后,所述对所述候选实体词中文本元素对应的文本表示进行整合处理,得到所述候选实体词的文本表示,包括:

根据文本序列顺序依次选取所述候选实体词中的文本元素;其中,所述文本序列顺序是从所述待识别文本中第一个文本元素依次到最后一个文本元素;

通过循环神经网络模型,对选取的文本元素的文本表示依次进行前向传播处理,并

将与所述候选实体词中最后一个文本元素对应的输出,确定为所述候选实体词的文本表示。

3.根据权利要求1所述的命名实体识别方法,其特征在于,在所述将每次滑动后所述卷积窗口所覆盖的文本元素组合为候选实体词之后,所述对所述候选实体词中文本元素对应的文本表示进行整合处理,得到所述候选实体词的文本表示,包括:

通过卷积神经网络模型,对所述候选实体词中文本元素对应的文本表示进行前向传播处理,得到所述候选实体词的文本表示;

其中,所述卷积神经网络模型的卷积核尺寸与所述卷积窗口的长度一致。

4.根据权利要求1所述的命名实体识别方法,其特征在于,所述对所述候选实体词的文本表示进行分类处理,以在候选类别中确定所述候选实体词所属的类别,包括:

对所述候选实体词的文本表示进行全连接处理；

通过第一分类函数对全连接处理后的所述候选实体词的文本表示进行映射处理，得到与多个所述候选类别一一对应的概率；

将数值最大的概率对应的候选类别，确定为所述候选实体词所属的类别；

其中，所述第一分类函数用于对所述候选实体词进行二分类。

5.根据权利要求1所述的命名实体识别方法，其特征在于，所述对所述候选实体词的文本表示进行分类处理，以在候选类别中确定所述候选实体词所属的类别，包括：

对所述候选实体词的文本表示进行全连接处理；

通过第二分类函数对全连接处理后的所述候选实体词的文本表示进行映射处理，得到与多个所述候选类别一一对应的概率；

将超过概率阈值的概率对应的候选类别，确定为所述候选实体词所属的类别；

其中，所述第二分类函数用于对所述候选实体词进行多分类。

6.根据权利要求1至5任一项所述的命名实体识别方法，其特征在于，还包括：

对所述待识别文本进行分割处理得到多个语句；

将属于命名实体类别的、且出现频率满足频率条件的候选实体词确定为摘要关键词；

根据所述语句包括的摘要关键词的数量，确定所述语句的评分；

将评分满足评分条件的语句，确定为所述待识别文本的文本摘要。

7.根据权利要求1至5任一项所述的命名实体识别方法，其特征在于，还包括：

当所述待识别文本用于表示待推荐对象时，将属于命名实体类别的候选实体词确定为关键词；

获取用户画像关键词，并确定所述用户画像关键词与所述待推荐对象对应的关键词之间的关键词重合度；

当所述关键词重合度超过第一重合度阈值时，执行推荐所述待推荐对象的操作。

8.根据权利要求7所述的命名实体识别方法，其特征在于，所述确定所述用户画像关键词与所述待推荐对象对应的关键词之间的关键词重合度，包括：

确定所述用户画像关键词与所述待推荐对象对应的关键词之间的交集，并确定所述交集包括的关键词的第一数量；

确定所述用户画像关键词与所述待推荐对象对应的关键词之间的并集，并确定所述并集包括的关键词的第二数量；

将所述第一数量与所述第二数量之间的比值，确定为所述用户画像关键词与所述待推荐对象对应的关键词之间的关键词重合度。

9.根据权利要求1至5任一项所述的命名实体识别方法，其特征在于，还包括：

将属于命名实体类别的候选实体词确定为关键词；

确定第一待识别文本与第二待识别文本之间的关键词重合度；

当所述关键词重合度超过第二重合度阈值时，将所述第一待识别文本与所述第二待识别文本划分为同一个文本类。

10.根据权利要求1至5任一项所述的命名实体识别方法，其特征在于，还包括：

将属于命名实体类别的候选实体词确定为关键词；

对所述待识别文本进行句法分析处理，得到所述待识别文本中的主语关键词、关系词

及宾语关键词;其中,所述关系词用于表示所述主语关键词与所述宾语关键词之间的关系;

根据所述主语关键词、所述关系词及所述宾语关键词构建三元组,并将所述三元组添加至知识图谱;

其中,所述知识图谱用于响应包括所述主语关键词及所述关系词的查询请求。

11.一种基于人工智能的命名实体识别装置,其特征在于,包括:

向量转换模块,用于对待识别文本中的文本元素进行向量转换处理,得到所述文本元素的文本表示;其中,所述文本元素的类型包括字和词;

遍历模块,用于根据文本跨度遍历所述待识别文本中的文本元素,以将总长度不超过所述文本跨度的文本元素组成候选实体词;

整合模块,用于对所述候选实体词中文本元素对应的文本表示进行整合处理,得到所述候选实体词的文本表示;

分类模块,用于对所述候选实体词的文本表示进行分类处理,以在候选类别中确定所述候选实体词所属的类别;其中,所述候选类别包括非实体类别和多个命名实体类别,其中,所述遍历模块具体用于执行以下任意一种处理:

对所述待识别文本中的文本元素进行第一层次遍历循环,其中,所述第一层次遍历循环包括多次第一层次遍历;将每个所述第一层次遍历得到的文本元素确定为起点元素;针对在每个所述第一层次遍历中确定的起点元素,执行包括多次第二层次遍历的第二层次遍历循环;根据在所述第二层次遍历循环中已经执行的第二层次遍历的次数,确定同步增大或缩小的扫描范围,根据所述扫描范围对所述待识别文本中的文本元素进行从所述起点元素开始的第二层次遍历,并将所述起点元素和所述第二层次遍历得到的文本元素组合为候选实体词,直至得到的候选实体词的长度等于所述文本跨度;或

根据多个不同长度的卷积窗口执行以下操作:在所述待识别文本中执行所述卷积窗口的滑动操作,且每次滑动操作的幅度为一个文本元素;将每次滑动后所述卷积窗口所覆盖的文本元素组合为候选实体词;其中,所述卷积窗口的长度小于或等于所述文本跨度。

12.一种电子设备,其特征在于,包括:

存储器,用于存储可执行指令;

处理器,用于执行所述存储器中存储的可执行指令时,实现权利要求1至10任一项所述的基于人工智能的命名实体识别方法。

13.一种计算机可读存储介质,其特征在于,存储有可执行指令,用于引起处理器执行时,实现权利要求1至10任一项所述的基于人工智能的命名实体识别方法。

基于人工智能的命名实体识别方法、装置及电子设备

技术领域

[0001] 本发明涉及人工智能技术，尤其涉及一种基于人工智能的命名实体识别方法、装置、电子设备及计算机可读存储介质。

背景技术

[0002] 人工智能(AI，Artificial Intelligence)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法和技术及应用系统。自然语言处理(NLP，Nature Language processing)是人工智能中的一个重要方向，主要研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。

[0003] 命名实体识别是自然语言处理的一个重要研究分支，旨在对文本中的命名实体进行定位并分类为预先定义的类别，命名实体如人员、组织、位置或数量等。在一些情况中，某个命名实体中的某一部分可能也是命名实体，即存在多层嵌套。对于多层嵌套的命名实体，在相关技术提供的方案中，通常是对字(token)进行多层标签标注，并通过编解码的方式来实现命名实体识别。但是，该种方式需要根据标签内容制定对应的解码规则，在标签数量较多时，解码规则将会变得难以设计和实现，进行命名实体识别的复杂度高。

发明内容

[0004] 本发明实施例提供一种基于人工智能的命名实体识别方法、装置、电子设备及计算机可读存储介质，能够提升命名实体识别的效率和精度。

[0005] 本发明实施例的技术方案是这样实现的：

[0006] 本发明实施例提供一种基于人工智能的命名实体识别方法，包括：

[0007] 对待识别文本中的文本元素进行向量转换处理，得到所述文本元素的文本表示；其中，所述文本元素的类型包括字和词；

[0008] 根据文本跨度遍历所述待识别文本中的文本元素，以将总长度不超过所述文本跨度的文本元素组成候选实体词；

[0009] 对所述候选实体词中文本元素对应的文本表示进行整合处理，得到所述候选实体词的文本表示；

[0010] 对所述候选实体词的文本表示进行分类处理，以在候选类别中确定所述候选实体词所属的类别；其中，所述候选类别包括非实体类别和多个命名实体类别。

[0011] 本发明实施例提供一种基于人工智能的命名实体识别装置，包括：

[0012] 向量转换模块，用于对待识别文本中的文本元素进行向量转换处理，得到所述文本元素的文本表示；其中，所述文本元素的类型包括字和词；

[0013] 遍历模块，用于根据文本跨度遍历所述待识别文本中的文本元素，以将总长度不超过所述文本跨度的文本元素组成候选实体词；

[0014] 整合模块，用于对所述候选实体词中文本元素对应的文本表示进行整合处理，得

到所述候选实体词的文本表示；

[0015] 分类模块，用于对所述候选实体词的文本表示进行分类处理，以在候选类别中确定所述候选实体词所属的类别；其中，所述候选类别包括非实体类别和多个命名实体类别。

[0016] 本发明实施例提供一种电子设备，包括：

[0017] 存储器，用于存储可执行指令；

[0018] 处理器，用于执行所述存储器中存储的可执行指令时，实现本发明实施例提供的基于人工智能的命名实体识别方法。

[0019] 本发明实施例提供一种计算机可读存储介质，存储有可执行指令，用于引起处理器执行时，实现本发明实施例提供的基于人工智能的命名实体识别方法。

[0020] 本发明实施例具有以下有益效果：

[0021] 本发明实施例通过遍历的方式，将总长度不超过文本跨度的文本元素组成候选实体词，并对候选实体词中文本元素对应的文本表示进行整合处理，得到候选实体词的文本表示，进而根据候选实体词的文本表示确定候选实体词所属的类别，提升了命名实体识别的效率和灵活性，也提升了确定出的类别的准确性。

附图说明

[0022] 图1是本发明实施例提供的基于人工智能的命名实体识别系统的一个可选的架构示意图；

[0023] 图2是本发明实施例提供的服务器的一个可选的架构示意图；

[0024] 图3是本发明实施例提供的基于人工智能的命名实体识别装置的一个可选的架构示意图；

[0025] 图4A是本发明实施例提供的基于人工智能的命名实体识别方法的一个可选的流程示意图；

[0026] 图4B是本发明实施例提供的基于人工智能的命名实体识别方法的一个可选的流程示意图；

[0027] 图4C是本发明实施例提供的基于人工智能的命名实体识别方法的一个可选的流程示意图；

[0028] 图4D是本发明实施例提供的基于人工智能的命名实体识别方法的一个可选的流程示意图；

[0029] 图5是本发明实施例提供的使用循环神经网络模型进行命名实体识别的一个可选的架构示意图；

[0030] 图6是本发明实施例提供的使用卷积神经网络模型进行命名实体识别的一个可选的架构示意图；

[0031] 图7是本发明实施例提供的进行问答的一个可选的流程示意图。

具体实施方式

[0032] 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进一步地详细描述，所描述的实施例不应视为对本发明的限制，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其它实施例，都属于本发明保护的范围。

[0033] 在以下的描述中,涉及到“一些实施例”,其描述了所有可能实施例的子集,但是可以理解,“一些实施例”可以是所有可能实施例的相同子集或不同子集,并且可以在不冲突的情况下相互结合。

[0034] 在以下的描述中,所涉及的术语“第一\第二”仅仅是是区别类似的对象,不代表针对对象的特定排序,可以理解地,“第一\第二”在允许的情况下可以互换特定的顺序或先后次序,以使这里描述的本发明实施例能够以除了在这里图示或描述的以外的顺序实施。

[0035] 除非另有定义,本文所使用的所有的技术和科学术语与属于本发明的技术领域的技术人员通常理解的含义相同。本文中所使用的术语只是为了描述本发明实施例的目的,不是旨在限制本发明。

[0036] 对本发明实施例进行进一步详细说明之前,对本发明实施例中涉及的名词和术语进行说明,本发明实施例中涉及的名词和术语适用于如下的解释。

[0037] 1)自然语言处理:是人工智能的一个重要方向,研究在人与人交际中以及人与计算机交际中的语言问题。在自然语言处理中,主要研制表示语言能力(linguistic competence)和语言应用(linguistic performance)的模型,建立计算框架来实现这样的语言模型,提出相应的方法来不断完善这样的语言模型,根据这样的语言模型设计各种实用系统,并探讨这些实用系统的评测技术。

[0038] 2)卷积神经网络(CNN,Convolutional Neural Network)模型:一种前馈神经网络模型,人工神经元可以响应周围单元,卷积神经网络模型通常包括卷积层和池化层。

[0039] 3)循环神经网络(RNN,Recurrent Neural Network)模型:一类用于处理序列数据的神经网络模型,这种模型的内部状态可以展示动态时序行为,可以利用内部的记忆来处理变长的输入序列。

[0040] 4)命名实体:文本中一些特殊的词或词组,根据实际应用场景的不同,这些词或词组也具有不同的类别和规定。常见的命名实体如人员、组织、位置、时间表达式、数量、货币值及百分比等。

[0041] 5)多重类别的命名实体:即同时属于多个命名实体类别的命名实体。

[0042] 6)多层嵌套的命名实体:某个命名实体中的某一部分也是命名实体,则称为多层嵌套。例如在文本“光荣的中国人民解放军”中,“中国人民解放军”是“军队”类别的实体,嵌套在命名实体“中国人民解放军”内的“中国”也是一个命名实体。

[0043] 7)序列标注(Sequence Tagging):通常是指对于一个线性输入序列,给序列中的每个元素打上标签集合中的某个标签的过程。

[0044] 8)命名实体识别(NER,Named Entity Recognition):旨在将文本中的命名实体定位并分类为预先定义的类别,如人员、组织及位置等。NER广泛应用于信息提取、问答系统、句法分析及机器翻译等应用场景。

[0045] 9)Nested NER:传统的NER通常只能从文本中粗糙地识别出平面结构的文本,不同于传统的NER,Nested NER的任务目标是从文本中识别出多重类别的命名实体以及多层嵌套的命名实体。

[0046] 10)文本表示:文字是人类认知过程中产生的高层认知抽象实体,在自然语言处理中,需要将文字转换为计算机可以处理的数据类型,即是转换为向量形式的文本表示。

[0047] 对于Nested NER任务,相关技术主要提供了Nested NE BILOU编解码方案来实现,

该方案主要对字(token)进行序列标注,并基于序列到序列的模型结构来进行编解码处理。在编码过程中,采用类似BILOU的标注编码方式,为了适应多层嵌套以及多重类别的情况,该方案允许多层的字(token)标签标注。在解码过程中,为了应对编码得到的多层标签,该方案通过制定一定的规则来处理这些多层标签结果,例如就近匹配等。作为示例,提供了如下所示的编码示例表:

[0048]

In	O
the	B-ORG
US	I-ORG|U-GPE
Federal	I-ORG
District	I-ORG|U-GPE
Court	I-ORG
of	I-ORG
New	I-ORG|B-GPE
Mexico	I-ORG|L-GPE
.	O

[0049] 其中,“B”表示命名实体的起始位置,“I”表示命名实体的中间,“L”表示命名实体的结尾,“O”表示不属于命名实体,“U”表示一个单独的命名实体,“ORG”表示组织机构,“GPE”表示地理政治实体。

[0050] 但是,该方案由于需要根据数据的标签内容制定好一定的解码规则,这就使得在命名实体类别的数量较多时,解码规则将会变得难以设计和实现,在Nested NER任务中,进行命名实体识别的效率和灵活性差。

[0051] 本发明实施例提供一种基于人工智能的命名实体识别方法、装置、电子设备及计算机可读存储介质,能够提升命名实体识别的效率和精度。

[0052] 下面说明本发明实施例提供的电子设备的示例性应用,本发明实施例提供的电子设备可以是服务器,例如部署在云端的服务器,根据获取到的待识别文本,向用户提供远程的命名实体识别功能;也可以是终端设备,例如问答设备,根据命名实体识别得到的命名实体扩充知识图谱,并根据知识图谱实现智能问答;甚至可以是手持终端等设备。

[0053] 参见图1,图1是本发明实施例提供的基于人工智能的命名实体识别系统100的一个可选的架构示意图,为实现支撑一个基于人工智能的命名实体识别应用,终端设备400(示例性示出了终端设备400-1和终端设备400-2)通过网络300连接服务器200,网络300可以是广域网或者局域网,又或者是二者的组合。

[0054] 在一些实施例中,终端设备400可在本地执行本发明实施例提供的基于人工智能的命名实体识别方法,具体获取用户录入或自动选择的待识别文本,通过遍历方式得到待识别文本中的候选实体词,并对候选实体词的文本表示进行分类处理,得到候选实体词所属的类别。

[0055] 除此之外,服务器200也可以执行本发明实施例提供的基于人工智能的命名实体识别方法,具体从终端设备400获取待识别文本,经一系列处理后,将识别出的属于命名实体类别的候选实体词发送至终端设备400,以使终端设备400的用户知悉。

[0056] 终端设备400可以在图形界面410(示例性示出了图形界面410-1和图形界面410-

2)中显示命名实体识别过程中的各种结果,例如属于命名实体类别的候选实体词等,在图1中以待识别文本“光荣的中国人民解放军”为例,示出了待识别文本中属于命名实体类别的候选实体词包括“中国”、“中国人”及“中国人民解放军”。命名实体识别的结果可应用于NLP领域的各个应用场景,例如图1所示的摘要确定、对象推荐、文本归类及问答系统的场景,又例如信息抽取、语法分析及机器翻译的场景,将在后文阐述命名实体的具体应用。

[0057] 下面继续说明本发明实施例提供的电子设备的示例性应用。电子设备可以实施为笔记本电脑,平板电脑,台式计算机,机顶盒,移动设备(例如,移动电话,便携式音乐播放器,个人数字助理,专用消息设备,便携式游戏设备)等各种类型的终端设备,也可以实施为服务器。

[0058] 下面,以电子设备为服务器为例进行说明。参见图2,图2是本发明实施例提供的服务器200(例如,可以是图1所示的服务器200)的架构示意图,图2所示的服务器200包括:至少一个处理器210、存储器240和至少一个网络接口220。服务器200中的各个组件通过总线系统230耦合在一起。可理解,总线系统230用于实现这些组件之间的连接通信。总线系统230除包括数据总线之外,还包括电源总线、控制总线和状态信号总线。但是为了清楚说明起见,在图2中将各种总线都标为总线系统230。

[0059] 处理器210可以是一种集成电路芯片,具有信号的处理能力,例如通用处理器、数字信号处理器(DSP,Digital Signal Processor),或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等,其中,通用处理器可以是微处理器或者任何常规的处理器等。

[0060] 存储器240可以是可移除的,不可移除的或其组合。示例性的硬件设备包括固态存储器,硬盘驱动器,光盘驱动器等。存储器240可选地包括在物理位置上远离处理器210的一个或多个存储设备。

[0061] 存储器240包括易失性存储器或非易失性存储器,也可包括易失性和非易失性存储器两者。非易失性存储器可以是只读存储器(ROM,Read Only Memory),易失性存储器可以是随机存取存储器(RAM,Random Access Memory)。本发明实施例描述的存储器240旨在包括任意适合类型的存储器。

[0062] 在一些实施例中,存储器240能够存储数据以支持各种操作,这些数据的示例包括程序、模块和数据结构或者其子集或超集,下面示例性说明。

[0063] 操作系统241,包括用于处理各种基本系统服务和执行硬件相关任务的系统程序,例如框架层、核心库层、驱动层等,用于实现各种基础业务以及处理基于硬件的任务;

[0064] 网络通信模块242,用于经由一个或多个(有线或无线)网络接口220到达其他计算设备,示例性的网络接口220包括:蓝牙、无线相容性认证(WiFi)、和通用串行总线(USB,Universal Serial Bus)等。

[0065] 在一些实施例中,本发明实施例提供的基于人工智能的命名实体识别装置可以采用软件方式实现,图2示出了存储在存储器240中的基于人工智能的命名实体识别装置243,其可以是程序和插件等形式的软件,包括以下软件模块:向量转换模块2431、遍历模块2432、整合模块2433及分类模块2434,这些模块是逻辑上的,因此根据所实现的功能可以进行任意的组合或进一步拆分。将在下文中说明各个模块的功能。

[0066] 在另一些实施例中,本发明实施例提供的基于人工智能的命名实体识别装置可以

采用硬件方式实现，作为示例，本发明实施例提供的基于人工智能的命名实体识别装置可以是采用硬件译码处理器形式的处理器，其被编程以执行本发明实施例提供的基于人工智能的命名实体识别方法，例如，硬件译码处理器形式的处理器可以采用一个或多个应用专用集成电路(ASIC，Application Specific Integrated Circuit)、DSP、可编程逻辑器件(PLD，Programmable Logic Device)、复杂可编程逻辑器件(CPLD，Complex Programmable Logic Device)、现场可编程门阵列(FPGA，Field-Programmable Gate Array)或其他电子元件。

[0067] 本发明实施例提供的基于人工智能的命名实体识别方法可以由上述的服务器执行，也可以由终端设备(例如，可以是图1所示的终端设备400-1和终端设备400-2)执行，或者由服务器和终端设备共同执行。

[0068] 下面将结合上文记载的电子设备的示例性应用和结构，说明电子设备中通过嵌入的基于人工智能的命名实体识别装置，而实现基于人工智能的命名实体识别方法的过程。

[0069] 参见图3和图4A，图3是本发明实施例提供的基于人工智能的命名实体识别装置243的架构示意图，示出了通过一系列模块实现命名实体识别的流程，图4A是本发明实施例提供的基于人工智能的命名实体识别方法的流程示意图，将结合图3对图4A示出的步骤进行说明。

[0070] 在步骤101中，对待识别文本中的文本元素进行向量转换处理，得到文本元素的文本表示；其中，文本元素的类型包括字和词。

[0071] 作为示例，参见图3，在向量转换模块2431，获取待识别文本，其中，待识别文本可以是由用户录入的，如通过语音、手写或其他方式录入，也可以是自动选择的。然后，确定待识别文本中的文本元素，文本元素的类型包括字和词。特别地，对于文本元素是词的情况，对待识别文本进行分词处理，得到文本元素，分词处理可通过语言技术平台(LTP，Language Technology Platform)工具或其他工具实现。值得说明的是，还可对待识别文本进行分割处理，例如以待识别文本中的逗号和句号为分割位置进行分割，得到待识别文本中的语句，再确定语句中的文本元素。

[0072] 对待识别文本中的每个文本元素进行向量转换处理，即对文本元素进行嵌入(Embedding)处理，将文本元素映射至向量空间，得到向量形式的文本表示，便于后续处理。

[0073] 在步骤102中，根据文本跨度遍历待识别文本中的文本元素，以将总长度不超过文本跨度的文本元素组成候选实体词。

[0074] 这里，总长度相当于文本元素的数量，即组成的候选实体词包括的文本元素的数量不超过文本跨度。通过遍历操作，可完备地枚举出待识别文本中可能存在的命名实体，适用于存在多层嵌套的命名实体的情况。值得说明的是，遍历操作中所用的文本跨度可由人为设定，如设定为7，也可获取数据库中包括的文本元素最多的命名实体，并将该命名实体包括的文本元素的数量确定为文本跨度。

[0075] 在步骤103中，对候选实体词中文本元素对应的文本表示进行整合处理，得到候选实体词的文本表示。

[0076] 经步骤102的遍历操作后，得到多个候选实体词，对于其中的每个候选实体词，对候选实体词中所有文本元素对应的文本表示进行整合处理，得到该候选实体词的文本表示。根据遍历操作的不同，整合处理的方式也随之不同，具体内容在后文进行详细阐述。

[0077] 在步骤104中，对候选实体词的文本表示进行分类处理，以在候选类别中确定候选实体词所属的类别；其中，候选类别包括非实体类别和多个命名实体类别。

[0078] 这里，对候选实体词的文本表示进行分类处理，将其映射为与多个候选类别一一对应的概率，并根据多个概率确定候选实体词所属的类别。值得说明的是，在本发明实施例中，将非实体同样作为一个候选类别，并与命名实体类别同等对待，即候选类别包括非实体类别和多个命名实体类别。另外，本发明实施例中的“多个”是指至少两个。

[0079] 在得到候选实体词所属的类别后，可将待识别文本中属于命名实体类别的候选实体词，应用于自然语言处理的应用场景，包括但不限于图3所示的摘要确定、对象推荐、文本归类及问答的应用场景。

[0080] 在一些实施例中，步骤104之后，还包括：对待识别文本进行分割处理得到多个语句；将属于命名实体类别的、且出现频率满足频率条件的候选实体词确定为摘要关键词；根据语句包括的摘要关键词的数量，确定语句的评分；将评分满足评分条件的语句，确定为待识别文本的文本摘要。

[0081] 在本发明实施例中，可根据命名实体识别的结果，来确定待识别文本的文本摘要，其中，待识别文本可以是论文、新闻或评论文章等，对此不做限定。首先，对待识别文本进行分割处理，例如以逗号和句号为分割位置进行分割，得到待识别文本中的多个语句，当然，分割处理的操作也可在步骤101之前进行，本发明实施例对此不做限定。

[0082] 为了便于区分，将待识别文本中属于命名实体类别的候选实体词命名为关键词，并将出现频率满足频率条件的关键词确定为摘要关键词，例如将出现频率最高的K个关键词确定为摘要关键词，其中K为大于0的整数，当然，也可将超过频率阈值的频率对应的关键词确定为摘要关键词，频率阈值可根据实际应用场景设定。

[0083] 对于待识别文本包括的每个语句，确定语句包括的摘要关键词的数量，并根据该数量确定语句的评分，例如，可直接将语句包括的摘要关键词的数量作为该语句的评分，也可将语句包括的摘要关键词的数量与该语句包括的文本元素的数量相除，得到该语句的评分。最后，将评分满足评分条件的语句，确定为待识别文本的文本摘要，例如可以将评分最高的L个语句，确定为待识别文本的文本摘要，其中L为大于0的整数，也可以将评分超过评分阈值的语句，确定为待识别文本的文本摘要。通过上述方式，提升了确定摘要的准确性，使得确定出的文本摘要能够较好地表示待识别文本的整体语义。

[0084] 在一些实施例中，步骤104之后，还包括：当待识别文本用于表示待推荐对象时，将属于命名实体类别的候选实体词确定为关键词；获取用户画像关键词，并确定用户画像关键词与待推荐对象对应的关键词之间的关键词重合度；当关键词重合度超过第一重合度阈值时，执行推荐待推荐对象的操作。

[0085] 这里，待识别文本用于表示待推荐对象，例如，待识别文本是某个商品或某部电影的描述文本。根据对待识别文本进行命名实体识别的结果，可实现针对性的智能对象推荐。具体地，将待识别文本中属于命名实体类别的候选实体词确定为关键词，同时，获取用户画像关键词，并确定用户画像关键词与待推荐对象对应的关键词之间的关键词重合度，其中，用户画像关键词可以是用户设定的，也可以是对用户的历史浏览记录进行关键词统计得到的。

[0086] 在确定关键词重合度时，可确定用户画像关键词与待推荐对象对应的关键词之间

的交集和并集，将交集包括的关键词的第一数量与并集包括的关键词的第二数量相除，得到关键词重合度。以待识别文本是某部电影的描述性文本举例，待识别文本中的关键词包括“爱情”、“文艺”和“剧情”，而用户画像关键词包括“爱情”、“科幻”和“喜剧”，则可得到关键词重合度为1/5。

[0087] 对于关键词重合度，通过设定第一重合度阈值判断是否进行推荐，具体地，当关键词重合度超过第一重合度阈值(如80％)时，执行推荐待推荐对象的操作。本发明实施例对推荐的具体方式不做限定，例如可以是邮件推荐、短信推荐或前端弹窗推荐等。通过上述方式，使得推荐的对象更加符合用户画像，增加了推荐的准确性。

[0088] 在一些实施例中，步骤104之后，还包括：将属于命名实体类别的候选实体词确定为关键词；确定第一待识别文本与第二待识别文本之间的关键词重合度；当关键词重合度超过第二重合度阈值时，将第一待识别文本与第二待识别文本划分为同一个文本类。

[0089] 命名实体识别的结果也可用于文本归类的应用场景，具体地，将属于命名实体类别的候选实体词确定为关键词，并确定第一待识别文本与第二待识别文本之间的关键词重合度。这里，在计算关键词重合度时，也可按照上述方式，确定第一待识别文本的关键词与第二待识别文本的关键词之间的交集和并集，将交集包括的关键词的数量与并集包括的关键词的数量相除，得到第一待识别文本与第二待识别文本之间的关键词重合度。

[0090] 在文本归类的应用场景中，还设定第二重合度阈值，当关键词重合度超过第二重合度阈值(如80％)时，证明第一待识别文本与第二待识别文本较为相似，将第一待识别文本与第二待识别文本划分为同一个文本类。划分为同一个文本类的文本可用于进行相似文本推荐，例如当查询请求的查询目标是第一待识别文本时，返回第一待识别文本以及与第一待识别文本属于同一文本类的第二待识别文本，以响应该查询请求。通过上述方式，将关键词重合度作为文本之间的相似度，进行文本归类，提升了归类的准确性，归类后的文本可用于相似文本推荐，即是将属于同一文本类的文本共同推荐至用户，提升了用户体验。

[0091] 在一些实施例中，步骤104之后，还包括：将属于命名实体类别的候选实体词确定为关键词；对待识别文本进行句法分析处理，得到待识别文本中的主语关键词、关系词及宾语关键词；其中，关系词用于表示主语关键词与宾语关键词之间的关系；根据主语关键词、关系词及宾语关键词构建三元组，并将三元组添加至知识图谱；其中，知识图谱用于响应包括主语关键词及关系词的查询请求。

[0092] 命名实体识别的结果可用于新词发现，例如，将待识别文本中属于命名实体类别的候选实体词确定为关键词，并将关键词及关键词所属的类别共同添加至知识图谱，从而实现命名实体的有效扩充。除此之外，还可将关键词之间的关系添加至知识图谱，具体地，基于已识别出的关键词，对待识别文本进行句法分析处理，例如可通过LTP工具或其他工具进行句法分析处理，得到待识别文本中的主语关键词、关系词及宾语关键词，其中，关系词用于表示主语关键词与宾语关键词之间的关系。例如，待识别文本为“张三向李四借钱”，则通过句法分析处理，可得到主语关键词为“张三”，关系词为“借钱”，宾语关键词为“李四”。

[0093] 根据得到的主语关键词、关系词及宾语关键词构建主语-谓语-宾语(SPO，Subject-Predication-Object)三元组，并将SPO三元组添加至知识图谱，成为知识图谱中的一条知识。知识图谱可用于响应包括主语关键词及关系词的查询请求，例如当查询请求用于查询张三借钱的对象时，可在知识图谱中进行查询，并根据结果“李四”进行应答。通过

上述方式，实现了命名实体以及命名实体之间关系的有效扩充，扩充后的知识图谱可用于进行更加精确的问答。

[0094] 通过发明实施例对于图4A的上述示例性实施可知，本发明实施例通过遍历的方式枚举待识别文本中的所有候选实体词，能够有效应对命名实体存在多层嵌套的情况，相较于相关技术提供的方案，能够大大提升命名实体识别的效率和灵活性。

[0095] 在一些实施例中，参见图4B，图4B是本发明实施例提供的基于人工智能的命名实体识别方法的一个可选的流程示意图，图4A示出的步骤102可以通过步骤201至步骤203实现，将结合各步骤进行说明。

[0096] 在步骤201中，对待识别文本中的文本元素进行第一层次遍历循环，其中，第一层次遍历循环包括多次第一层次遍历。

[0097] 作为示例，参见图3，在遍历模块2432中，可应用第一层次遍历和第二层次遍历的方式，来得到候选实体词。首先，对待识别文本中的文本元素进行第一层次遍历循环，其中，第一层次遍历循环包括多次第一层次遍历。本发明实施例对第一层次遍历的遍历顺序不做限定，例如遍历顺序可以是从待识别文本中第一个文本元素依次到最后一个文本元素，也可以是从待识别文本中最后一个文本元素依次到第一个文本元素。

[0098] 在步骤202中，将每个第一层次遍历得到的文本元素确定为起点元素。

[0099] 这里，为了便于区分，将每个第一层次遍历得到的文本元素，确定为起点元素，该起点元素为候选实体词的起点，具体为候选实体词的词首或词尾。

[0100] 在步骤203中，针对在每个第一层次遍历中确定的起点元素，执行包括多次第二层次遍历的第二层次遍历循环，以将起点元素和第二层次遍历得到的文本元素组合为候选实体词。

[0101] 针对每一个确定出的起点元素，执行包括多次第二层次遍历的第二层次遍历循环。对于第二层次遍历循环中的每一次第二层次遍历，根据在第二层次遍历循环中已经执行的第二层次遍历的次数，确定同步增大或缩小的扫描范围，根据扫描范围对待识别文本中的文本元素进行从起点元素开始的第二层次遍历，并将起点元素和第二层次遍历得到的文本元素组合为候选实体词，直至得到的候选实体词的长度等于文本跨度，确定第二层次遍历循环执行完成。其中，扫描范围在确定起点元素时初始化，扫描范围每次增大或缩小的单位为一个文本元素，另外，候选实体词的长度是指候选实体词包括的文本元素的数量。值得说明的是，本发明实施例对第二层次遍历的遍历顺序同样不做限定。

[0102] 为了便于理解，以第一层次遍历和第二层次遍历的遍历顺序均为从待识别文本中第一个文本元素依次到最后一个文本元素进行说明，且文本跨度为7，待识别文本为“光荣的中国人民解放军”，文本元素为字，则在第一层次遍历循环中，第一次第一层次遍历得到的文本元素为“光”，将“光”确定为起点元素，进入第二层次遍历循环。在第二层次遍历循环中，以扫描范围初始化为0，且扫描范围与已经执行的第二层次遍历的次数同步增大进行举例，则在第一次第二层次遍历中，由于扫描范围为0，故得到的候选实体词为“光”；在第二次第二层次遍历中，扫描范围增大为1，得到的候选实体词为“光荣”；在第三次第二层次遍历中，扫描范围增大为2，得到的候选实体词为“光荣的”，以此类推，直到得到候选实体词“光荣的中国人民”，停止第二层次遍历循环。然后，根据下一次第一层次遍历得到的文本元素，即以“荣”为起点元素开始新的第二层次遍历循环。

[0103] 在图4B中，图4A示出的步骤103可以通过步骤204至步骤205实现，将结合各步骤进行说明。

[0104] 在步骤204中，根据文本序列顺序依次选取候选实体词中的文本元素；其中，文本序列顺序是从待识别文本中第一个文本元素依次到最后一个文本元素。

[0105] 这里，根据从待识别文本中第一个文本元素依次到最后一个文本元素的顺序，选取步骤203得到的候选实体词中的文本元素。例如候选实体词为“光荣的”，则依次选取“光”、“荣”和“的”。

[0106] 在步骤205中，通过循环神经网络模型，对选取的文本元素的文本表示依次进行前向传播处理，并将与候选实体词中最后一个文本元素对应的输出，确定为候选实体词的文本表示。

[0107] 作为示例，参见图3，在整合模块2433中，针对于包括第一层次遍历和第二层次遍历的遍历方式，通过RNN模型来实现序列数据的处理。具体地，将步骤204中逐一选取的文本元素所对应的文本表示，依次按照逐个时刻(step)输入至RNN模型中，并将RNN模型的与候选实体词中最后一个文本元素对应时刻的输出，作为该候选实体词的文本表示。由于RNN模型的序列记忆性能及语义表示能力较强，故得到的候选实体词的文本表示较为准确。

[0108] 通过发明实施例对于图4B的上述示例性实施可知，本发明实施例通过第一次遍历及第二次遍历的方式构建候选实体词，实现了候选实体词的有效枚举，能够适用于多层嵌套的情况，并且，通过适于处理序列数据的RNN模型进行整合处理，能够实现候选实体词的文本表示准确性的显著提升。

[0109] 在一些实施例中，参见图4C，图4C是本发明实施例提供的基于人工智能的命名实体识别方法的一个可选的流程示意图，图4A示出的步骤102可更新为步骤301，在步骤301中，根据多个不同长度的卷积窗口执行以下操作：在待识别文本中执行卷积窗口的滑动操作，且每次滑动操作的幅度为一个文本元素；将每次滑动后卷积窗口所覆盖的文本元素组合为候选实体词。

[0110] 作为示例，参见图3，在遍历模块2432中，针对文本跨度设置多个不同长度的卷积窗口，且每个卷积窗口的长度均小于或等于文本跨度。为了提升遍历的完整性，可从长度为1的卷积窗口开始递增长度，直至得到长度达到文本跨度的卷积窗口，如文本跨度为7，则设置长度依次为1、2、……、7的卷积窗口。

[0111] 对于每个卷积窗口，在待识别文本中执行卷积窗口的滑动操作，且每次滑动操作的幅度为一个文本元素，这里，对执行滑动操作的顺序不做限定，例如可以是从待识别文本中第一个文本元素到最后一个文本元素，也可以是从待识别文本中最后一个文本元素到第一个文本元素。在每次滑动后，将卷积窗口所覆盖的文本元素组合为候选实体词。

[0112] 举例来说，待识别文本为“光荣的中国人民解放军”，某个卷积窗口的长度为3，且以待识别文本中第一个文本元素到最后一个文本元素的顺序进行滑动，则在第一次滑动时，得到候选实体词“光荣的”；在第二次滑动时，得到候选实体词“荣的中”，以此类推。

[0113] 在图4C中，图4A示出的步骤103可更新为步骤302，在步骤302中，通过卷积神经网络模型，对候选实体词中文本元素对应的文本表示进行前向传播处理，得到候选实体词的文本表示；其中，卷积神经网络模型的卷积核尺寸与卷积窗口的长度一致。

[0114] 作为示例，参见图3，在整合模块2433中，对于每个卷积窗口，将得到的候选实体词

中文本元素对应的文本表示输入至CNN模型中，将CNN模型经前向传播处理得到的输出，确定为候选实体词的文本表示，其中，该CNN模型的卷积核尺寸(size)与卷积窗口的长度一致。

[0115] 通过发明实施例对于图4C的上述示例性实施可知，本发明实施例通过滑动卷积窗口的方式构建候选实体词，从另一个角度实现了候选实体词的有效枚举，并且，与卷积窗口对应的CNN模型进行整合处理，提升了得到的候选实体词的文本表示的准确性。

[0116] 在一些实施例中，参见图4D，图4D是本发明实施例提供的基于人工智能的命名实体识别方法的一个可选的流程示意图，图4A示出的步骤104可以通过步骤401至步骤405实现，将结合各步骤进行说明。

[0117] 在步骤401中，对候选实体词的文本表示进行全连接处理。

[0118] 作为示例，参见图3，在分类模块2434中，通过全连接层对候选实体词的文本表示进行全连接处理，便于后续进行分类。

[0119] 在步骤402中，通过第一分类函数对全连接处理后的候选实体词的文本表示进行映射处理，得到与多个候选类别一一对应的概率。

[0120] 作为示例，参见图3，在分类模型2434中，当命名实体识别的任务是二分类任务时，通过第一分类函数对全连接处理后的候选实体词的文本表示进行映射处理，得到与多个候选类别一一对应的概率，其中，第一分类函数可以是Softmax分类函数。值得说明的是，本发明实施例中的二分类是指针对不存在多重类别的命名实体的情况，判断候选实体词是属于非实体类别还是命名实体类别。

[0121] 在步骤403中，将数值最大的概率对应的候选类别，确定为候选实体词所属的类别。

[0122] 这里，将数值最大的概率对应的候选类别，确定为候选实体词所属的类别，候选实体词所属的类别为非实体类别或命名实体类别。

[0123] 在步骤404中，通过第二分类函数对全连接处理后的候选实体词的文本表示进行映射处理，得到与多个候选类别一一对应的概率。

[0124] 作为示例，参见图3，在分类模型2434中，当命名实体识别的任务是多分类任务时，通过第二分类函数对全连接处理后的候选实体词的文本表示进行映射处理，得到与多个候选类别一一对应的概率，其中，第二分类函数可以是Sigmoid分类函数。值得说明的是，本发明实施例中的多分类是指针对存在多重类别的命名实体的情况，判断候选实体词是属于非实体类别，还是属于至少一个命名实体类别。

[0125] 在步骤405中，将超过概率阈值的概率对应的候选类别，确定为候选实体词所属的类别。

[0126] 在多分类的情况下，设定概率阈值，并将超过概率阈值的概率对应的候选类别，确定为候选实体词所属的类别，此时，候选实体词所属的类别可能是一个或多个。

[0127] 通过发明实施例对于图4D的上述示例性实施可知，本发明实施例对于二分类和多分类的情况，使用不同的分类函数进行分类处理，提升了命名实体识别的灵活性。

[0128] 下面，将说明本发明实施例在实际的应用场景中的示例性应用。

[0129] 本发明实施例提供了如图5所示的使用循环神经网络模型进行命名实体识别的一个可选的架构示意图，下面按照由底向上的顺序进行依次解释。

[0130] 1)文本输入模块&文本表示模块。

[0131] 该模块泛指具备文本表示能力的NLP模型结构以及基于表示结构的其他扩展,可根据实际应用场景进行设定,例如该模块可应用BERT模型及其改进等。将待识别文本输入至该模块,得到待识别文本中文本元素的文本表示,其中,文本元素为字或词,在图5中以文本元素为字的情况进行说明。

[0132] 2)对应文本表示。

[0133] 在图5中,待识别文本为“光荣的中国人民解放军”,经文本输入模块&文本表示模块处理后,得到的待识别文本中每个字的向量形式的文本表示,在图5中以方框为示例。

[0134] 3)循环神经网络模型。

[0135] 本发明实施例中的RNN模型泛指RNN模型及其变种,例如长短时记忆网络(LSTM,Long Short-Term Memory)模型及门控循环单元(GRU,Gated Recurrent Unit)模型等。图5中以双向RNN模型举例,模型中的每一个圆表示某一个时刻的细胞(cell)。

[0136] 对于构建候选实体词的过程,可划分为两层循环,分别是第一层次遍历循环和第二层次遍历循环。在第一层次遍历循环中,按照待识别文本中第一个文本元素到最后一个文本元素的顺序进行多次第一层次遍历,第一层次遍历循环的目的是枚举并完整覆盖待识别文本中所有可能的候选实体词的词首。以上述的待识别文本举例来说,第一层次遍历循环所遍历的候选实体词的词首依次为:

[0137] 光->荣->的->中->国->人->民->解->放->军

[0138] 将第一层次遍历循环中的每一次第一层次遍历得到的词首对应的文本表示,作为RNN模型中第一时刻(step)的输入向量。值得说明的是,对于上层遍历循环中的每一次第一层次遍历,其对RNN模型的输入均彼此独立进行,即前一次第一层次遍历得到的词首对应的文本表示,将不再作为后一次的RNN模型的输入,后一次第一层次遍历得到的词首的文本表示,将重新使用RNN模型,即作为RNN模型的第一时刻的输入向量。

[0139] 在每一次第一层次遍历内,即确定了词首之后,执行一次完整的第二层次遍历循环。第二层次遍历循环是指,以第一层次遍历得到的词首为候选实体词的开始,根据逐渐增大的扫描范围,循环扫描以该词首为开始的每一个候选实体词,直到得到的候选实体词包括的文本元素的数量达到文本跨度。举例来说,在第一层次遍历循环中,经第一层次遍历得到的词首为“中”,则开始执行第二层次遍历循环,即以“中”这个字为候选实体词的词首,逐一扩大扫描范围。以文本跨度为7举例,那么在第二层次遍历循环中,首先将“中”作为一个候选实体词,再增大1个单位(图5中示出的单位为字)的扫描范围,得到候选实体词“中国”,扫描范围再度增大,得到候选实体词“中国人”,以此类推,最终扫描至达到文本跨度的候选实体词,即“中国人民解放军”为止,以此为一个第二层次遍历循环。

[0140] 在组建候选实体词的同时,将逐一扫描到的字对应的文本表示依次按照逐个时刻输入至RNN模型中,并且以当前扫描到的字所对应的RNN时刻的输出,作为从“词首”到当前扫描字所构成的候选实体词的文本表示。举例来说,对于候选实体词“中国人”,将“中”、“国”和“人”对应的文本表示依次按照逐个时刻输入至RNN模型中,并将RNN模型的与“人”对应时刻的输出,确定为“中国人”的文本表示。通过上述的第一层次遍历循环和第二层次遍历循环,能够较好地地利用RNN模型的序列记忆性能,根据RNN模型较强的语义表示能力,能够得到较为准确的候选实体词的文本表示。

[0141] 4)候选实体词

[0142] 图5中示出了经遍历后,得到的不同的候选实体词,其与RNN模型输出的文本表示对应。

[0143] 5)Softmax/Sigmoid分类层

[0144] 该分类层用于对RNN模型输出的大量候选实体词的文本表示进行分类处理。本发明实施例提供了两种分类方式,分别使用Softmax和Sigmoid两种激活(分类)函数。当待识别文本中不存在多重类别的命名实体,即为二分类任务时,优先采用Softmax激活函数,将得到的每一个候选实体词的文本表示,使用全连接层等结构结合Softmax激活函数进行分类,最后得出候选实体词最有可能所属的一个类别。当待识别文本中存在多重类别的实体,即为多分类任务时,优先采用Sigmoid激活函数,同样使用全连接层等结构结合Sigmoid激活函数进行分类,最后通过设置概率阈值的方式,得出候选实体词可能所属的类别。值得说明的是,在分类过程中,非实体类别可与其他的命名实体类别平等看待。

[0145] 举例来说,如图5所示,经分类层进行分类处理后,得到候选实体词“中”所属的类别是非实体类别,候选实体词“中国”所属的类别是国家名称类别,候选实体词“中国人”所属的类别是群体类别,候选实体词“中国人民解放军”所属的类别是军队类别。完成命名实体识别后,可将得到的命名实体(即属于命名实体类别的候选实体词)应用于NLP领域的各种应用场景,例如在新词发现的场景中,可将得到的命名实体添加至知识图谱中,实现新的命名实体的自动挖掘。

[0146] 本发明实施例还提供了如图6所示的使用卷积神经网络模型进行命名实体识别的一个可选的架构示意图,下面按照由底向上的顺序进行依次解释。

[0147] 1)文本输入模块&文本表示模块。

[0148] 2)对应文本表示。

[0149] 图6中的1)和2)与图5对应的内容类似,在此不做赘述。

[0150] 3)CNN模型。

[0151] 此部分的CNN模型泛指CNN模型及其变种等。图6中以原始的CNN模型进行举例,针对预设的文本跨度,使用多种卷积核尺寸(size)的CNN模型,对候选实体词中文本元素的文本表示进行一维CNN处理。

[0152] 举例来说,在文本跨度为7的情况下,设置7种尺寸的卷积核,尺寸依次为1至7,卷积核的数量可根据具体实际应用场景进行设定,不同尺寸的卷积核的数量可设置为相同。同时,设置长度与卷积核尺寸相同的卷积窗口,并根据文本序列顺序对卷积窗口执行滑动操作,文本序列顺序为待识别文本中第一个文本元素依次到最后一个文本元素的顺序。将每次滑动后,卷积窗口所覆盖的文本元素组合为候选实体词,并将候选实体词中文本元素对应的文本表示,输入至拥有对应尺寸的卷积核的CNN模型,CNN模型的卷积输出结果即为候选实体词的文本表示。

[0153] 如图6所示,当长度为3的卷积窗口覆盖在“中国人”上时,将此时刻CNN模型(具有尺寸为3的卷积核)处理所得到的输出结果,作为“中国人”这个候选实体词的文本表示。

[0154] 4)候选实体词

[0155] 将卷积窗口所覆盖的文本元素组合为候选实体词,如图6所示的“中”、“中国”及“中国人”等候选实体词。

[0156] 5)Softmax/Sigmoid分类层

[0157] 图6中的5)与图5对应的内容类似,同样可针对不同的分类任务采用不同的激活函数,在此不做赘述。

[0158] 相较于相关技术提供的Nested NE BILOU编解码方案,通过本发明实施例提供的基于人工智能的命名实体识别方法,无需人为制定复杂的编解码规则,提升了命名实体识别的效率和灵活性。并且相较于相关技术提供的其他方案,本发明实施例也具有一定优势。

[0159] 具体地,相较于用于进行命名实体识别,且采用复杂结构的深度学习模型的MGNER方案,本发明实施例的模型结构更为简单,同时能够更加有效地增强对于候选实体词的语义表达能力,改善模型的性能指标。经发明人实验验证,在基于语言模型嵌入(ELMo,Embedding from Language Models)的相同文本表示结构下进行测试,两个方案的指标对比如下:

[0160] 在开源的ACE2004公开数据集下,MGNER模型方案的F1分数为79.5%,本发明实施例提供方法的F1分数为83.7%;在ACE2005公开数据集下,MGNER模型方案的F1分数为78.2%,本发明实施例提供方法的F1分数为82.4%,其中,F1分数为精确率和召回率的调和平均。

[0161] 另外,在相关技术提供的方案中,还可通过机器阅读理解(MRC,Machine Reading Comprehension)的思路来实现命名实体识别,但该方案针对每一种命名实体类别,均需要独立运行一次模型,所花费时间较长。而通过本发明实施例提供的方法,一次运行即可产出待识别文本中所有可能的候选实体词及其所属的类别,效率较高。使用具有115种命名实体类型的ACL-NNE公开数据集进行测试,并且在与MRC框架方案同样基于BERT模型的情况下,相较于MRC框架方案,本发明实施例提供的方法大约能够节省超过90%的时间。

[0162] 命名实体识别的结果可应用于NLP领域的各个应用场景,例如摘要确定、对象推荐、文本归类及问答系统的场景,又例如信息抽取、语法分析及机器翻译的场景,这里以问答系统的场景进行详细说明。

[0163] 本发明实施例提供了如图7所示的进行问答的流程示意图,在图7中,终端设备400-1和终端设备400-2由不同的用户持有,为了便于区分,将终端设备400-1命名为第一终端设备,将终端设备400-2命名为第二终端设备,将结合图7示出的各个步骤说明问答的过程。

[0164] 在步骤501中,第一终端设备将待识别文本发送至服务器。

[0165] 第一终端设备的用户通过第一终端设备,将待识别文本发送至服务器,这里对待识别文本的来源不做限定,例如待识别文本可以是某个人物的词条文本,或某个商品的说明文本等。

[0166] 在步骤502中,服务器通过遍历方式确定待识别文本中的候选实体词,并确定候选实体词所属的类别。

[0167] 这里,服务器通过遍历方式枚举待识别文本中所有可能出现的候选实体词,并确定每个候选实体词所属的类别,该过程与步骤101～步骤104类似,在此不做赘述。

[0168] 在步骤503中,服务器将属于命名实体类别的候选实体词确定为关键词,并对待识别文本进行句法分析处理,得到主语关键词、关系词及宾语关键词。

[0169] 在识别出属于命名实体类别的候选实体词后,服务器将该种候选实体词确定为关

键词，并进一步对待识别文本进行句法分析处理，得到存在依存关系的主语关键词、关系词及宾语关键词，其中，关系词用于表示主语关键词与宾语关键词之间的关系。例如，待识别文本为“张三向李四借钱”，则通过句法分析处理，得到主语关键词为“张三”，关系词为“借钱”，宾语关键词为“李四”。

[0170] 在步骤504，服务器根据主语关键词、关系词及宾语关键词构建三元组，并将三元组添加至知识图谱。

[0171] 这里，根据主语关键词、关系词及宾语关键词构建主-谓-宾三元组，例如主-谓-宾三元组为“张三-借钱-李四”。然后，服务器将构建的主-谓-宾三元组添加至知识图谱，同时，“张三”和“李四”也可作为命名实体添加在知识图谱中。

[0172] 在步骤505中，第二终端设备将查询请求发送至服务器。

[0173] 例如，第二终端设备将“张三向谁借钱?”的查询请求发送服务器。

[0174] 在步骤506中，服务器通过知识图谱确定查询请求的语义，并根据语义在知识图谱中进行查询，得到查询结果。

[0175] 这里，服务器可以将查询请求与知识图谱中的命名实体进行匹配，例如将“张三向谁借钱?”与知识图谱中的命名实体进行匹配，得到查询请求中与知识图谱匹配的命名实体为“张三”。除此之外，服务器也可以应用步骤101～步骤104的方式，得到查询请求中的候选实体词所属的类别，并将属于命名实体类别的候选实体词与知识图谱进行匹配。

[0176] 然后，服务器根据查询请求中与知识图谱匹配的命名实体，进一步进行句法分析处理，得到查询请求的语义，例如“张三-借钱-?”。服务器根据查询请求的语义，在知识图谱中进行查询，得到相应的查询结果，并将查询结果发送至第二终端设备，完成问答，例如将查询结果“李四”发送至第二终端设备。

[0177] 通过发明实施例对于图7的上述示例性实施可知，本发明实施例通过遍历的方式得到准确的命名实体识别结果，并根据命名实体识别结果进行知识图谱的扩充，提升了知识图谱中知识的准确性，也提升了基于知识图谱的问答系统的精度，使得用户在参与问答能够得到良好的用户体验。

[0178] 下面继续说明本发明实施例提供的基于人工智能的命名实体识别装置243实施为软件模块的示例性结构，在一些实施例中，如图2所示，存储在存储器240的基于人工智能的命名实体识别装置243中的软件模块可以包括：向量转换模块2431，用于对待识别文本中的文本元素进行向量转换处理，得到文本元素的文本表示；其中，文本元素的类型包括字和词；遍历模块2432，用于根据文本跨度遍历待识别文本中的文本元素，以将总长度不超过文本跨度的文本元素组成候选实体词；整合模块2433，用于对候选实体词中文本元素对应的文本表示进行整合处理，得到候选实体词的文本表示；分类模块2434，用于对候选实体词的文本表示进行分类处理，以在候选类别中确定候选实体词所属的类别；其中，候选类别包括非实体类别和多个命名实体类别。

[0179] 在一些实施例中，遍历模块2432，还用于：对待识别文本中的文本元素进行第一层次遍历循环，其中，第一层次遍历循环包括多次第一层次遍历；将每个第一层次遍历得到的文本元素确定为起点元素；针对在每个第一层次遍历中确定的起点元素，执行包括多次第二层次遍历的第二层次遍历循环：根据在第二层次遍历循环中已经执行的第二层次遍历的次数，确定同步增大或缩小的扫描范围，根据扫描范围对待识别文本中的文本元素进行从

起点元素开始的第二层次遍历，并将起点元素和第二层次遍历得到的文本元素组合为候选实体词，直至得到的候选实体词的长度等于文本跨度。

[0180] 在一些实施例中，整合模块2433，还用于：根据文本序列顺序依次选取候选实体词中的文本元素；其中，文本序列顺序是从待识别文本中第一个文本元素依次到最后一个文本元素；通过循环神经网络模型，对选取的文本元素的文本表示依次进行前向传播处理，并将与候选实体词中最后一个文本元素对应的输出，确定为候选实体词的文本表示。

[0181] 在一些实施例中，遍历模块2432，还用于：根据多个不同长度的卷积窗口执行以下操作：在待识别文本中执行卷积窗口的滑动操作，且每次滑动操作的幅度为一个文本元素；将每次滑动后卷积窗口所覆盖的文本元素组合为候选实体词；其中，卷积窗口的长度小于或等于文本跨度。

[0182] 在一些实施例中，整合模块2433，还用于：通过卷积神经网络模型，对候选实体词中文本元素对应的文本表示进行前向传播处理，得到候选实体词的文本表示；其中，卷积神经网络模型的卷积核尺寸与卷积窗口的长度一致。

[0183] 在一些实施例中，分类模块2434，还用于：对候选实体词的文本表示进行全连接处理；通过第一分类函数对全连接处理后的候选实体词的文本表示进行映射处理，得到与多个候选类别一一对应的概率；将数值最大的概率对应的候选类别，确定为候选实体词所属的类别；其中，第一分类函数用于对候选实体词进行二分类。

[0184] 在一些实施例中，分类模块2434，还用于：对候选实体词的文本表示进行全连接处理；通过第二分类函数对全连接处理后的候选实体词的文本表示进行映射处理，得到与多个候选类别一一对应的概率；将超过概率阈值的概率对应的候选类别，确定为候选实体词所属的类别；其中，第二分类函数用于对候选实体词进行多分类。

[0185] 在一些实施例中，基于人工智能的命名实体识别装置243还包括：分割模块，用于对待识别文本进行分割处理得到多个语句；摘要关键词确定模块，用于将属于命名实体类别的、且出现频率满足频率条件的候选实体词确定为摘要关键词；评分确定模块，用于根据语句包括的摘要关键词的数量，确定语句的评分；摘要确定模块，用于将评分满足评分条件的语句，确定为待识别文本的文本摘要。

[0186] 在一些实施例中，基于人工智能的命名实体识别装置243还包括：第一关键词确定模块，用于当待识别文本用于表示待推荐对象时，将属于命名实体类别的候选实体词确定为关键词；用户画像获取模块，用于获取用户画像关键词，并确定用户画像关键词与待推荐对象对应的关键词之间的关键词重合度；推荐模块，用于当关键词重合度超过第一重合度阈值时，执行推荐待推荐对象的操作。

[0187] 在一些实施例中，用户画像获取模块，还用于：确定用户画像关键词与待推荐对象对应的关键词之间的交集，并确定交集包括的关键词的第一数量；确定用户画像关键词与待推荐对象对应的关键词之间的并集，并确定并集包括的关键词的第二数量；将第一数量与第二数量之间的比值，确定为用户画像关键词与待推荐对象对应的关键词之间的关键词重合度。

[0188] 在一些实施例中，基于人工智能的命名实体识别装置243还包括：第二关键词确定模块，用于将属于命名实体类别的候选实体词确定为关键词；重合度计算模块，用于确定第一待识别文本与第二待识别文本之间的关键词重合度；归类模块，用于当关键词重合度超

过第二重合度阈值时，将第一待识别文本与第二待识别文本划分为同一个文本类。

[0189] 在一些实施例中，基于人工智能的命名实体识别装置243还包括：第三关键词确定模块，用于将属于命名实体类别的候选实体词确定为关键词；句法分析模块，用于对待识别文本进行句法分析处理，得到待识别文本中的主语关键词、关系词及宾语关键词；其中，关系词用于表示主语关键词与宾语关键词之间的关系；添加模块，用于根据主语关键词、关系词及宾语关键词构建三元组，并将三元组添加至知识图谱；其中，知识图谱用于响应包括主语关键词及关系词的查询请求。

[0190] 本发明实施例提供一种存储有可执行指令的计算机可读存储介质，其中存储有可执行指令，当可执行指令被处理器执行时，将引起处理器执行本发明实施例提供的方法，例如，如图4A、图4B、图4C或图4D示出的基于人工智能的命名实体识别方法。值得说明的是，计算机包括终端设备和服务器在内的各种计算设备。

[0191] 在一些实施例中，计算机可读存储介质可以是FRAM、ROM、PROM、EPROM、EEPROM、闪存、磁表面存储器、光盘、或CD-ROM等存储器；也可以是包括上述存储器之一或任意组合的各种设备。

[0192] 在一些实施例中，可执行指令可以采用程序、软件、软件模块、脚本或代码的形式，按任意形式的编程语言(包括编译或解释语言，或者声明性或过程性语言)来编写，并且其可按任意形式部署，包括被部署为独立的程序或者被部署为模块、组件、子例程或者适合在计算环境中使用的其它单元。

[0193] 作为示例，可执行指令可以但不一定对应于文件系统中的文件，可以可被存储在保存其它程序或数据的文件的一部分，例如，存储在超文本标记语言(HTML，Hyper Text Markup Language)文档中的一个或多个脚本中，存储在专用于所讨论的程序的单个文件中，或者，存储在多个协同文件(例如，存储一个或多个模块、子程序或代码部分的文件)中。

[0194] 作为示例，可执行指令可被部署为在一个计算设备上执行，或者在位于一个地点的多个计算设备上执行，又或者，在分布在多个地点且通过通信网络互连的多个计算设备上执行。

[0195] 综上，通过本发明实施例能够实现以下技术效果：

[0196] 1)本发明实施例提供了一种精炼且易用的候选实体词抽取方式，按照特定顺序逐一地、不同跨度地整合待识别文本中的序列信息，能够完备地枚举出待识别文本中的候选实体词。本发明实施例的模型结构简单，灵活性强，便于根据实际应用场景中的需要进行进一步改进，同时也易于移植到更多的深度学习模型中。

[0197] 2)本发明实施例契合了RNN模型及CNN模型在NLP领域中的应用特点，按照特定顺序，使用RNN模型或CNN模型的相关结构，对待识别文本中文本元素的文本表示进行整合处理，得到候选实体词的文本表示，能够更加简单有效地增强对于候选实体词的语义表达能力，改善模型的性能指标，兼顾简洁性和有效性。

[0198] 3)通过本发明实施例，一次运行即可产出待识别文本中所有的候选实体词及其所属的类别，能够节省较多时间，较快地得到候选实体词所属的类别。

[0199] 4)本发明实施例针对具体的分类任务(二分类/多分类)，采用对应的分类函数进行分类处理，提升了对于不同应用场景的适用性。

[0200] 5)在待识别文本包括多个语句的情况下，通过本发明实施例，能够根据属于命名

实体类别的候选实体词，确定每个语句的重要程度，从而筛选出文本摘要，提升了摘要选取的准确性。

[0201] 6)在对象推荐的场景下，本发明实施例通过将待识别文本中的关键词与用户画像关键词进行匹配，从而尽量推荐符合用户喜好的对象，提升了用户体验，也提升了推荐的对象的转化率。

[0202] 7)本发明实施例通过比对两个文本之间的关键词，根据得到的关键词重合度判断是否将两个文本归为一类，提升了文本归类的精度。

[0203] 8)本发明实施例在进行了命名实体识别后，可将属于命名实体类别的候选实体词添加至知识图谱，提升新词发现的准确度。除此之外，还可将待识别文本中出现的命名实体之间的关系添加至知识图谱，使得扩充后的知识图谱能够更好地应用于问答等场景。

[0204] 以上，仅为本发明的实施例而已，并非用于限定本发明的保护范围。凡在本发明的精神和范围之内所作的任何修改、等同替换和改进等，均包含在本发明的保护范围之内。

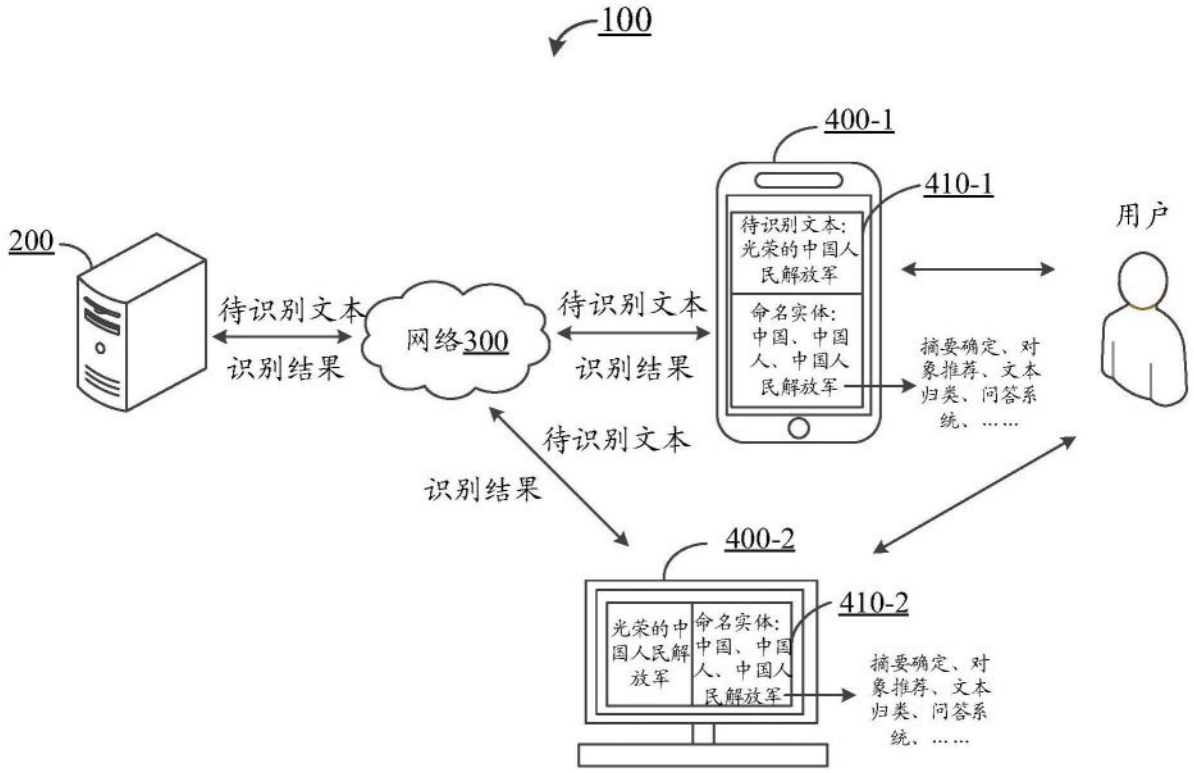

100
400-1
410-1
待识别文本:
光荣的中国人民解放军
命名实体:
中国、中国人、中国人民解放军
用户
200
待识别文本
识别结果
网络300
待识别文本
识别结果
摘要确定、对象推荐、文本归类、问答系统、……
待识别文本
识别结果
400-2
410-2
光荣的中国人民解放军
命名实体:
中国、中国人、中国人民解放军
摘要确定、对象推荐、文本归类、问答系统、……

图1

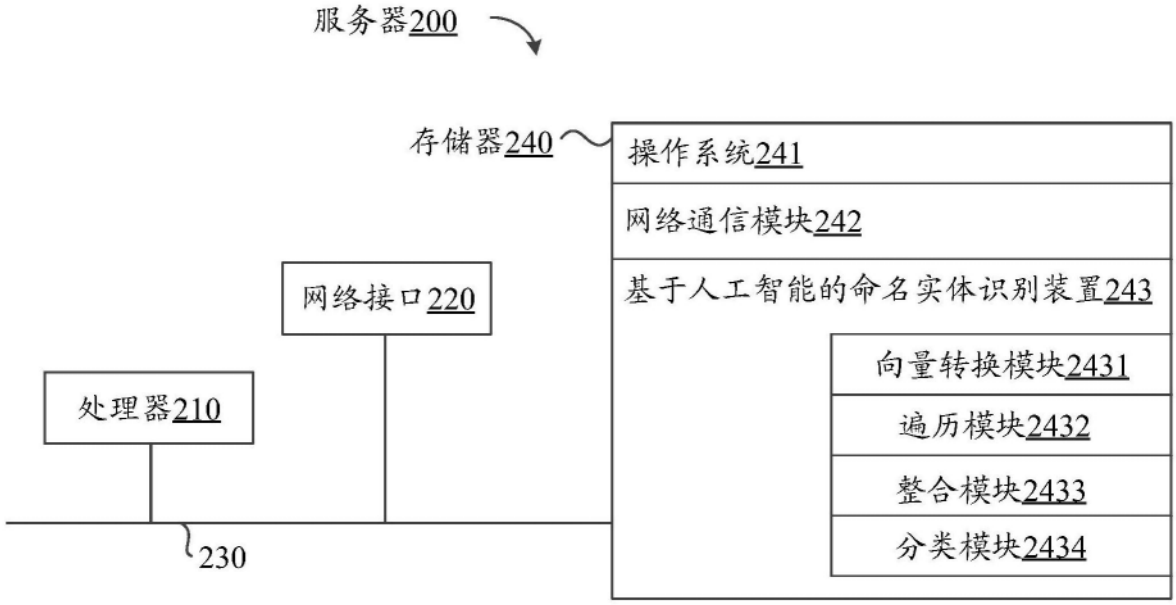

服务器200
存储器240
操作系统241
网络通信模块242
基于人工智能的命名实体识别装置243
向量转换模块2431
遍历模块2432
整合模块2433
分类模块2434
网络接口220
处理器210
230

图2

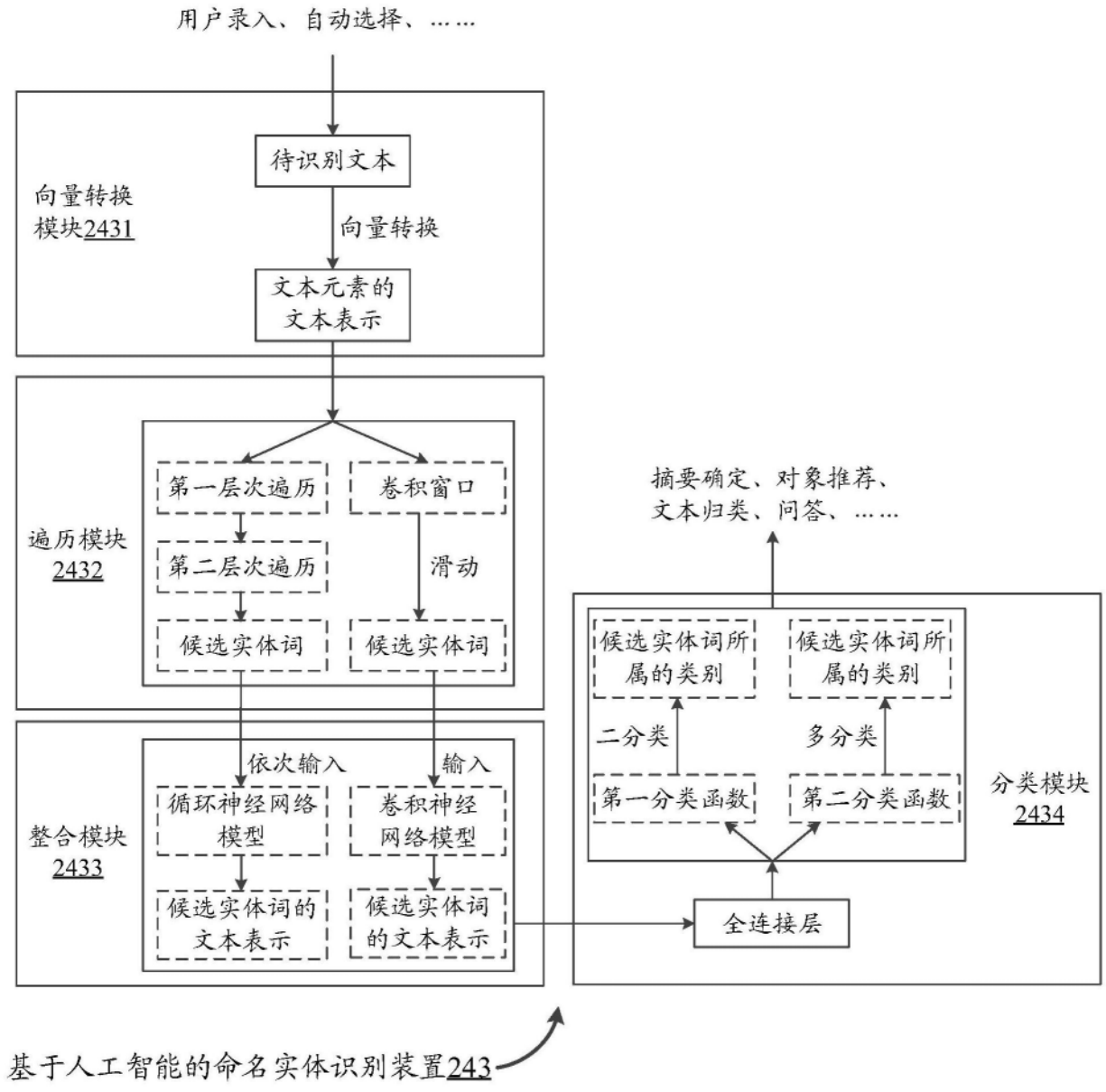

用户录入、自动选择、……
向量转换模块2431
待识别文本
向量转换
文本元素的文本表示
遍历模块2432
第一层次遍历
第二层次遍历
候选实体词
卷积窗口
滑动
候选实体词
整合模块2433
依次输入
输入
循环神经网络模型
卷积神经网络模型
候选实体词的文本表示
候选实体词的文本表示
摘要确定、对象推荐、文本归类、问答、……
候选实体词所属的类别
候选实体词所属的类别
二分类
多分类
第一分类函数
第二分类函数
分类模块2434
全连接层
基于人工智能的命名实体识别装置243

图3

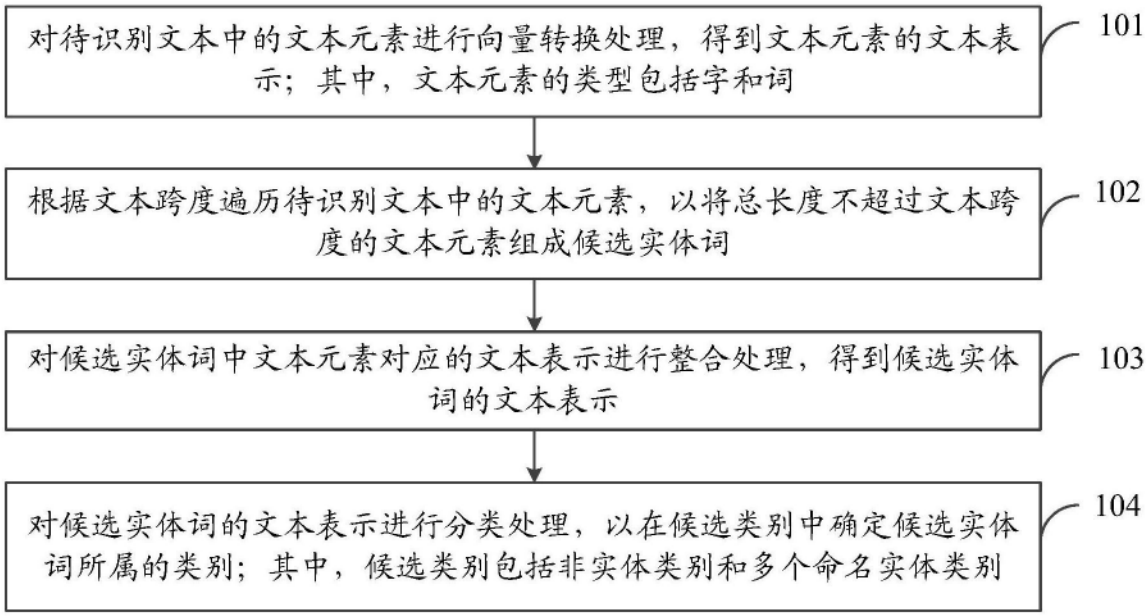

对待识别文本中的文本元素进行向量转换处理，得到文本元素的文本表示；其中，文本元素的类型包括字和词
101
根据文本跨度遍历待识别文本中的文本元素，以将总长度不超过文本跨度的文本元素组成候选实体词
102
对候选实体词中文本元素对应的文本表示进行整合处理，得到候选实体词的文本表示
103
对候选实体词的文本表示进行分类处理，以在候选类别中确定候选实体词所属的类别；其中，候选类别包括非实体类别和多个命名实体类别
104

图4A

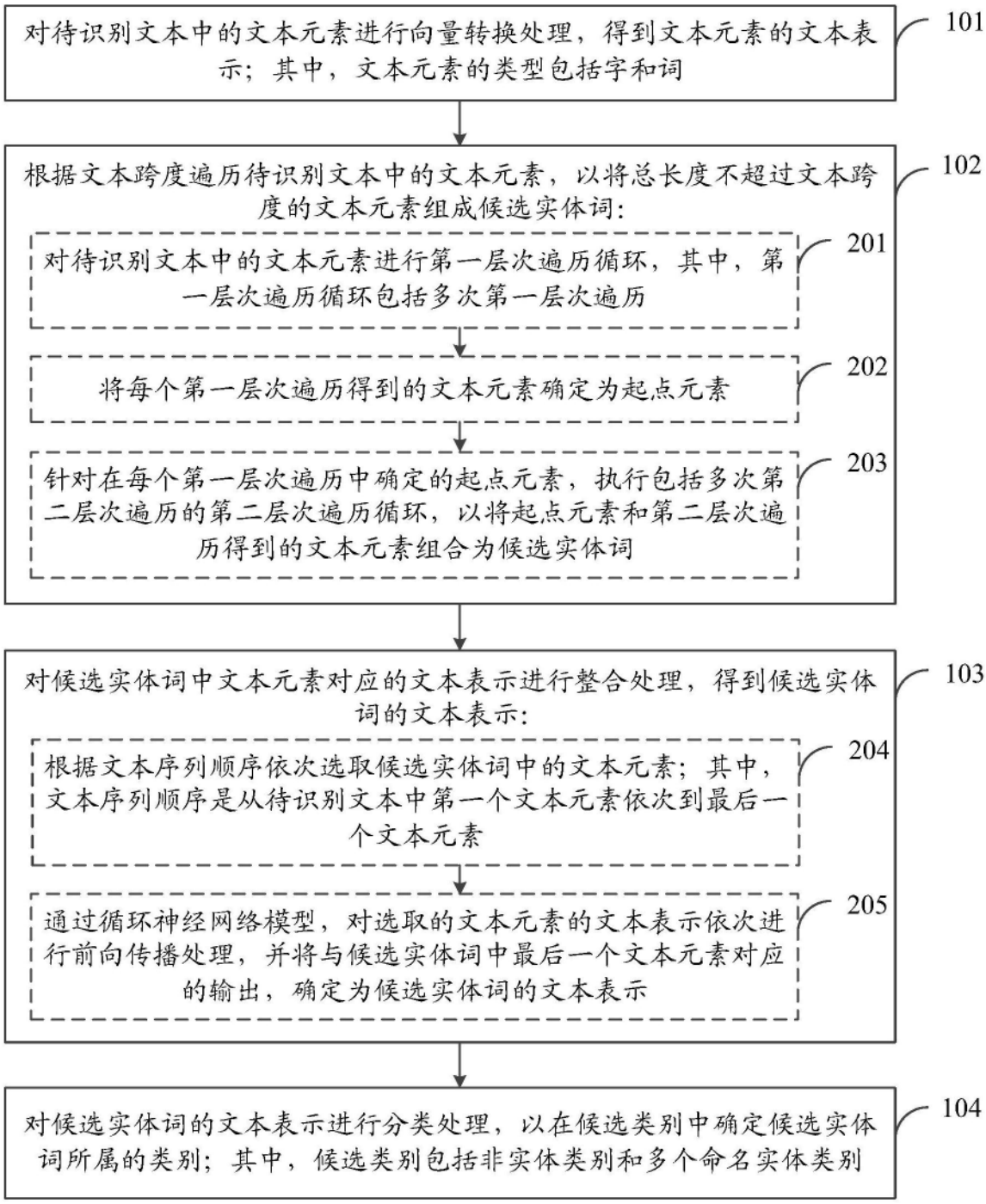

对待识别文本中的文本元素进行向量转换处理，得到文本元素的文本表示；其中，文本元素的类型包括字和词
101
根据文本跨度遍历待识别文本中的文本元素，以将总长度不超过文本跨度的文本元素组成候选实体词：
102
对待识别文本中的文本元素进行第一层次遍历循环，其中，第一层次遍历循环包括多次第一层次遍历
201
将每个第一层次遍历得到的文本元素确定为起点元素
202
针对在每个第一层次遍历中确定的起点元素，执行包括多次第二层次遍历的第二层次遍历循环，以将起点元素和第二层次遍历得到的文本元素组合为候选实体词
203
对候选实体词中文本元素对应的文本表示进行整合处理，得到候选实体词的文本表示：
103
根据文本序列顺序依次选取候选实体词中的文本元素；其中，文本序列顺序是从待识别文本中第一个文本元素依次到最后一个文本元素
204
通过循环神经网络模型，对选取的文本元素的文本表示依次进行前向传播处理，并将与候选实体词中最后一个文本元素对应的输出，确定为候选实体词的文本表示
205
对候选实体词的文本表示进行分类处理，以在候选类别中确定候选实体词所属的类别；其中，候选类别包括非实体类别和多个命名实体类别
104

图4B

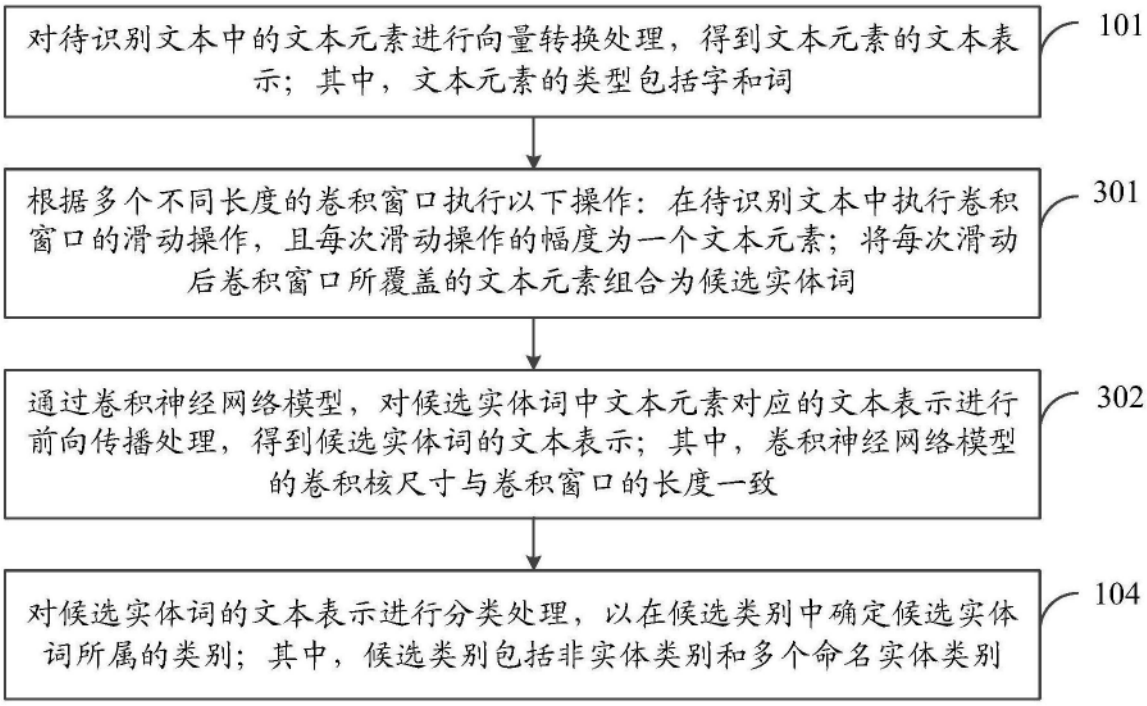

对待识别文本中的文本元素进行向量转换处理，得到文本元素的文本表示；其中，文本元素的类型包括字和词
101
根据多个不同长度的卷积窗口执行以下操作：在待识别文本中执行卷积窗口的滑动操作，且每次滑动操作的幅度为一个文本元素；将每次滑动后卷积窗口所覆盖的文本元素组合为候选实体词
301
通过卷积神经网络模型，对候选实体词中文本元素对应的文本表示进行前向传播处理，得到候选实体词的文本表示；其中，卷积神经网络模型的卷积核尺寸与卷积窗口的长度一致
302
对候选实体词的文本表示进行分类处理，以在候选类别中确定候选实体词所属的类别；其中，候选类别包括非实体类别和多个命名实体类别
104

图4C

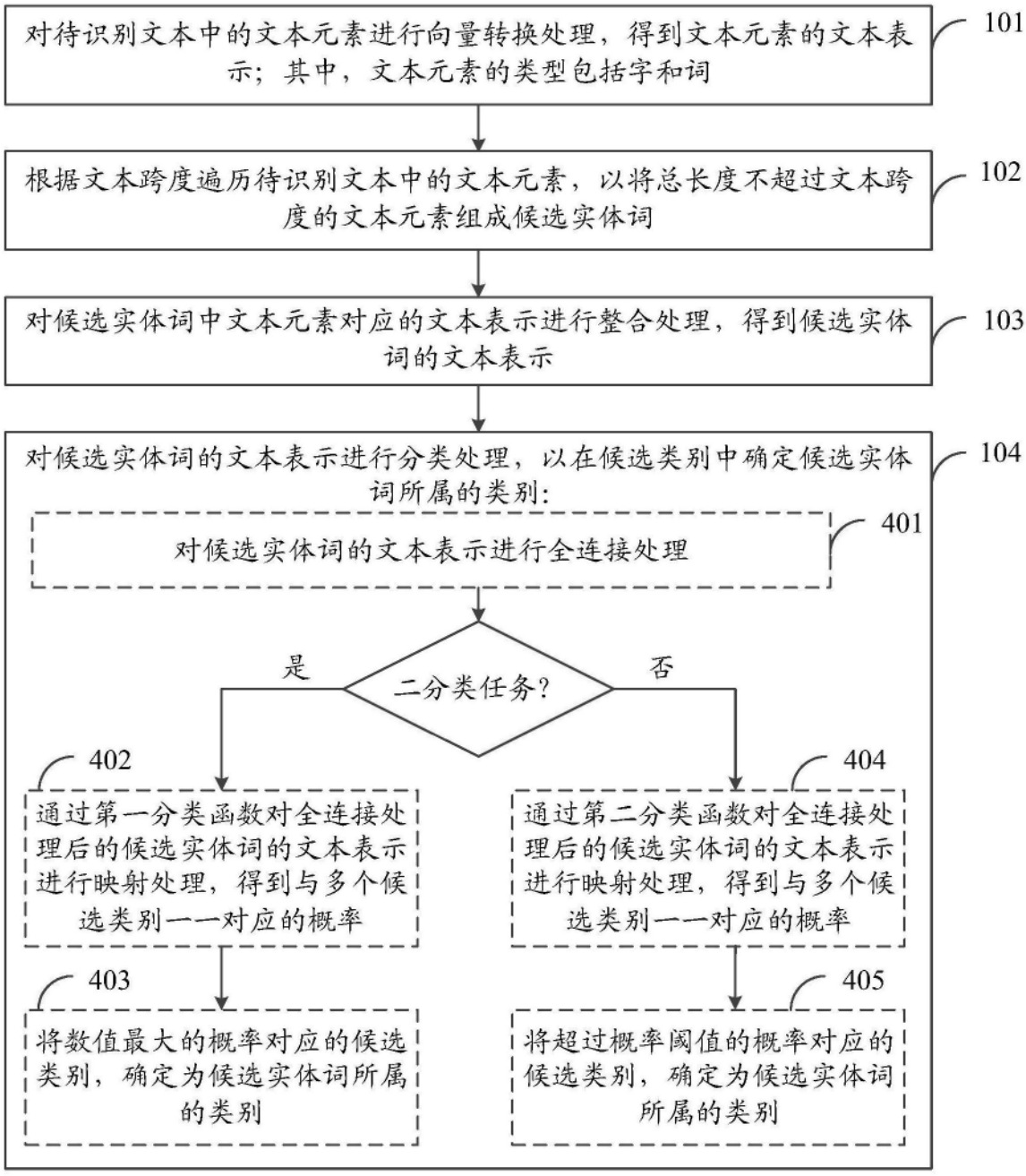

对待识别文本中的文本元素进行向量转换处理，得到文本元素的文本表示；其中，文本元素的类型包括字和词
101
根据文本跨度遍历待识别文本中的文本元素，以将总长度不超过文本跨度的文本元素组成候选实体词
102
对候选实体词中文本元素对应的文本表示进行整合处理，得到候选实体词的文本表示
103
对候选实体词的文本表示进行分类处理，以在候选类别中确定候选实体词所属的类别：
104
对候选实体词的文本表示进行全连接处理
401
二分类任务?
是
否
402
通过第一分类函数对全连接处理后的候选实体词的文本表示进行映射处理，得到与多个候选类别一一对应的概率
403
将数值最大的概率对应的候选类别，确定为候选实体词所属的类别
404
通过第二分类函数对全连接处理后的候选实体词的文本表示进行映射处理，得到与多个候选类别一一对应的概率
405
将超过概率阈值的概率对应的候选类别，确定为候选实体词所属的类别

图4D

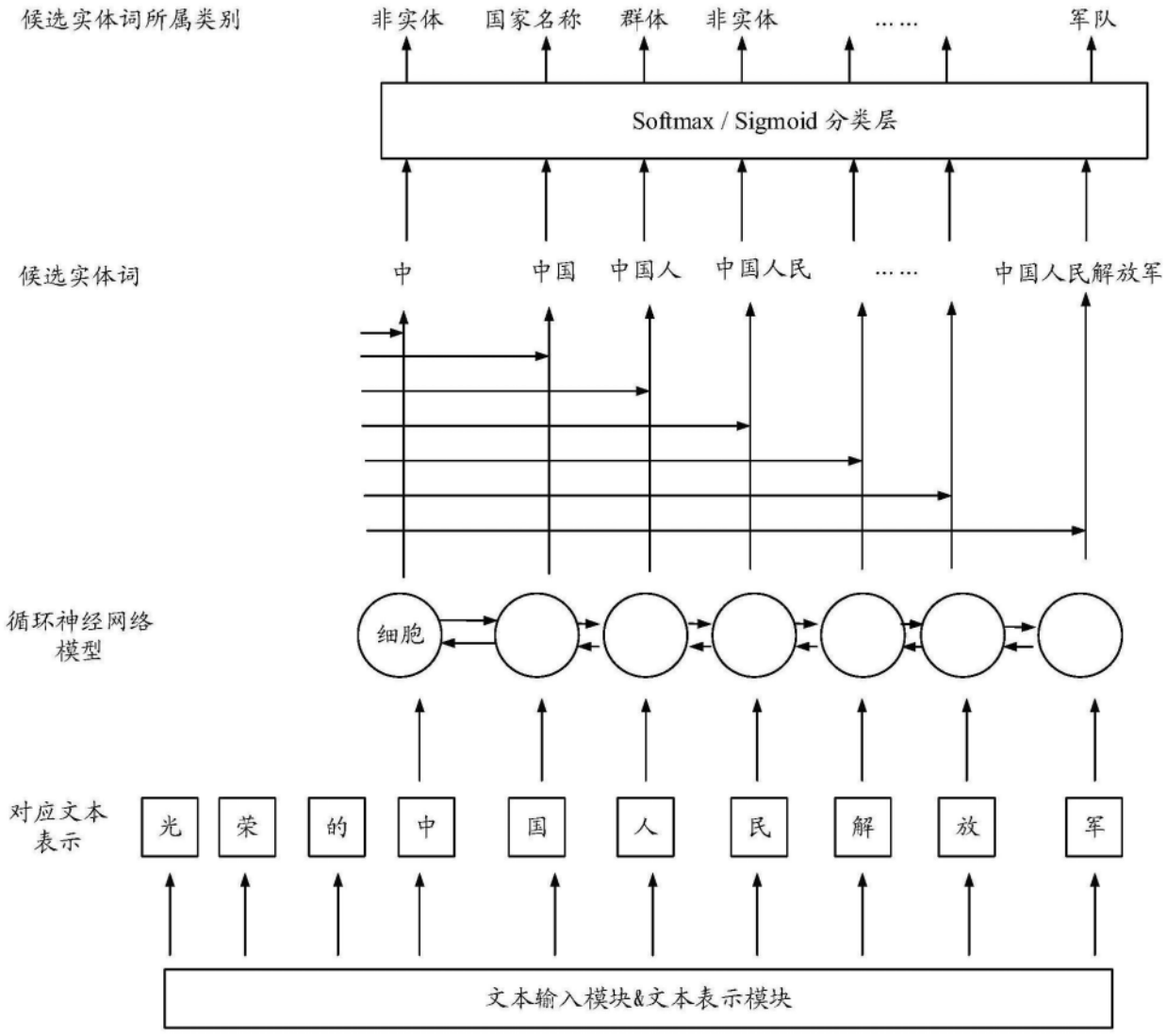

候选实体词所属类别
非实体
国家名称
群体
非实体
……
军队
Softmax / Sigmoid 分类层
候选实体词
中
中国
中国人
中国人民
……
中国人民解放军
循环神经网络模型
细胞
对应文本表示
光
荣
的
中
国
人
民
解
放
军
文本输入模块&文本表示模块

图5

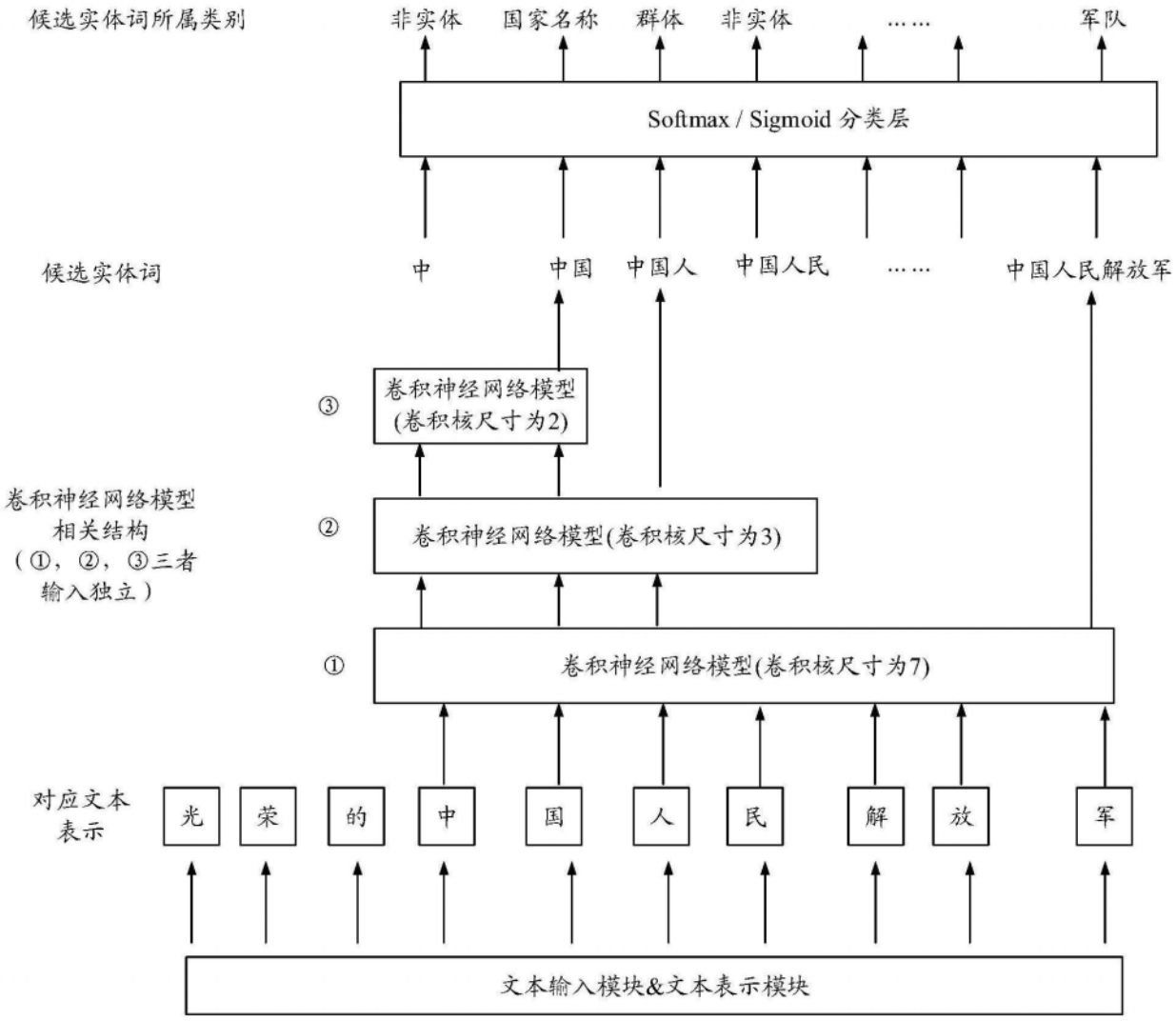

候选实体词所属类别
非实体
国家名称
群体
非实体
……
军队
Softmax / Sigmoid 分类层
候选实体词
中
中国
中国人
中国人民
……
中国人民解放军
③
卷积神经网络模型
(卷积核尺寸为2)
卷积神经网络模型
相关结构
（①，②，③三者
输入独立）
②
卷积神经网络模型(卷积核尺寸为3)
①
卷积神经网络模型(卷积核尺寸为7)
对应文本
表示
光
荣
的
中
国
人
民
解
放
军
文本输入模块&文本表示模块

图6

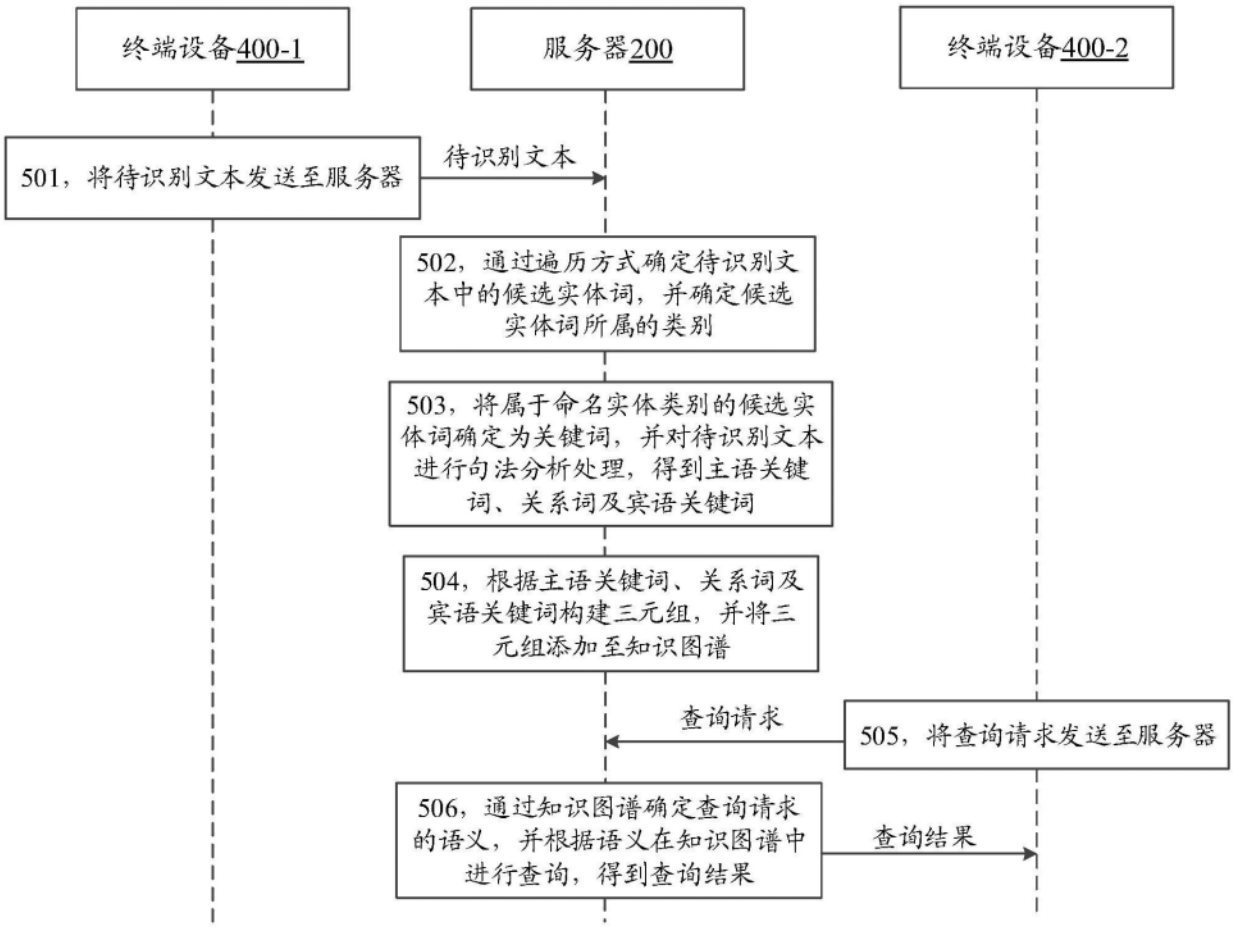

终端设备400-1
服务器200
终端设备400-2
501，将待识别文本发送至服务器
待识别文本
502，通过遍历方式确定待识别文本中的候选实体词，并确定候选实体词所属的类别
503，将属于命名实体类别的候选实体词确定为关键词，并对待识别文本进行句法分析处理，得到主语关键词、关系词及宾语关键词
504，根据主语关键词、关系词及宾语关键词构建三元组，并将三元组添加至知识图谱
查询请求
505，将查询请求发送至服务器
506，通过知识图谱确定查询请求的语义，并根据语义在知识图谱中进行查询，得到查询结果
查询结果

图7