(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110825875 B
(45)授权公告日 2022.12.06

(21)申请号 201911060988.X

(22)申请日 2019.11.01

(65)同一申请的已公布的文献号
申请公布号 CN 110825875 A

(43)申请公布日 2020.02.21

(73)专利权人 科大讯飞股份有限公司
地址 230088 安徽省合肥市高新开发区望江西路666号

(72)发明人 詹文超 沙晶 付瑞吉 王士进 魏思

(74)专利代理机构 北京路浩知识产权代理有限公司 11002
专利代理师 程琛

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/36(2019.01)

(56)对比文件
CN 109885825 A,2019.06.14
CN 107977361 A,2018.05.01
CN 109918680 A,2019.06.21
CN 110134954 A,2019.08.16
CN 104965819 A,2015.10.07
US 2019156212 A1,2019.05.23
US 2007174041 A1,2007.07.26
US 2019220749 A1,2019.07.18
CN 109271632 A,2019.01.25
武文雅等.基于高层语义注意力机制的中文实体关系抽取.《广西师范大学学报》.2019,第37卷(第1期),

审查员 单娟

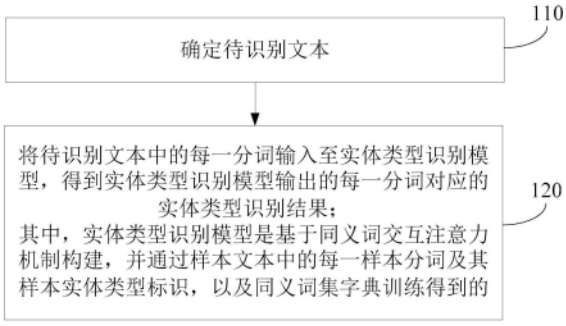
权利要求书3页 说明书11页 附图4页

(54)发明名称

文本实体类型识别方法、装置、电子设备和存储介质

(57)摘要

本发明实施例提供一种文本实体类型识别方法、装置、电子设备和存储介质，其中方法包括：确定待识别文本；将待识别文本中的每一分词输入至实体类型识别模型，得到实体类型识别模型输出的每一分词对应的实体类型识别结果；其中，实体类型识别模型是基于同义词交互注意力机制构建，并通过样本文本中的每一样本分词及其样本实体类型标识，以及同义词集字典训练得到的。本发明实施例提供的方法、装置、电子设备和存储介质，通过将待识别文本的每一分词输入至基于同义词交互注意力机制构建的实体类型识别模型中，进行实体类型识别，克服了由于文本实体表达方式的多变导致的识别困难的问题，有助于提高文本实体类型识别的准确性和可靠性。



CN 110825875 B

1.一种文本实体类型识别方法,其特征在于,包括:

确定待识别文本;

将所述待识别文本中的每一分词输入至实体类型识别模型,得到所述实体类型识别模型输出的每一分词对应的实体类型识别结果;

其中,所述实体类型识别模型是基于同义词交互注意力机制构建,并通过样本文本中的每一样本分词及其样本实体类型标识,以及同义词集字典训练得到的;

所述实体类型识别模型包括输入层、同义词交互注意力层和分类输出层;

对应地,所述将所述待识别文本中的每一分词输入至实体类型识别模型,得到所述实体类型识别模型输出的每一分词对应的实体类型识别结果,具体包括:

将所述待识别文本中的每一分词输入至所述输入层,得到所述输入层输出的每一分词的词向量;

将每一分词的词向量输入至所述同义词交互注意力层,得到所述同义词交互注意力层输出的每一分词的增强词向量;

将每一分词的增强词向量输入至所述分类输出层,得到所述分类输出层输出的每一分词的实体类型识别结果;

其中,所述同义词交互注意力层基于同义词集字典,判断输入的每一分词中是否存在同义词,从而对同义词的词向量进行交互增强,并输出每一分词的增强词向量。

2.根据权利要求1所述的文本实体类型识别方法,其特征在于,所述将每一分词的增强词向量输入至所述分类输出层,得到所述分类输出层输出的每一分词的实体类型识别结果,之前还包括:

针对任一分词,将所述任一分词的增强词向量更新为所述任一分词的增强词向量和所述任一分词的字符特征向量的拼接向量。

3.根据权利要求1所述的文本实体类型识别方法,其特征在于,所述实体类型识别模型还包括公式语义预测层;

对应地,所述将每一分词的词向量输入至所述同义词交互注意力层,得到所述同义词交互注意力层输出的每一分词的增强词向量,之后还包括:

基于每一分词的增强词向量,确定所述待识别文本中公式的词向量;

将所述公式的词向量输入至所述公式语义预测层,得到所述公式语义预测层输出的公式语义;

对应地,所述实体类型识别模型是基于样本文本中的每一样本分词及其样本实体类型标识,同义词集字典,以及所述样本文本中样本公式的样本公式语义训练得到的。

4.根据权利要求1所述的文本实体类型识别方法,其特征在于,所述将每一分词的词向量输入至所述同义词交互注意力层,得到所述同义词交互注意力层输出的每一分词的增强词向量,具体包括:

基于所述同义词集字典,从每一分词的词向量中选取任一分词的同义词的词向量,并构建所述任一分词的同义词集;所述同义词集包括每一同义词的词向量;

基于所述任一分词的词向量与所述同义词集中任一同义词的词向量,确定所述任一分词与所述任一同义词的相似度;

基于所述任一分词分别与所述同义词集中每一同义词的相似度,输出所述任一分词的

增强词向量。

5.根据权利要求4所述的文本实体类型识别方法,其特征在于,所述基于所述任一分词分别与所述同义词集中每一同义词的相似度,输出所述任一分词的增强词向量,具体包括:

基于所述任一分词分别与所述同义词集中每一同义词的相似度,确定每一同义词的词向量对应的权重;

基于所述每一同义词的词向量及其对应的权重,确定注意力词向量;

基于所述任一分词的词向量和所述注意力词向量,确定所述任一分词的增强词向量。

6.根据权利要求1所述的文本实体类型识别方法,其特征在于,所述分类输出层包括上下文层和分类层;

对应地,所述将每一分词的增强词向量输入至所述分类输出层,得到所述分类输出层输出的每一分词的实体类型识别结果,具体包括:

将每一分词的增强词向量输入至所述上下文层,得到所述上下文层输出的每一分词的序列向量;

将每一分词的序列向量输入至所述分类层,得到所述分类层输出的每一分词的实体类型识别结果。

7.根据权利要求6所述的文本实体类型识别方法,其特征在于,所述分类层包括softmax输出层和随机条件场CRF层;

对应地,所述将每一分词的序列向量输入至所述分类层,得到所述分类层输出的每一分词的实体类型识别结果,具体包括:

将每一分词的序列向量输入至所述softmax输出层,得到所述softmax输出层输出的每一分词的候选识别结果;

将每一分词的候选分类结果输入至所述随机条件场CRF层,得到所述随机条件场CRF层输出的每一分词的实体类型识别结果。

8.根据权利要求1至7中任一项所述的文本实体类型识别方法,其特征在于,所述待识别文本为数学文本。

9.一种文本实体类型识别装置,其特征在于,包括:

文本确定单元,用于确定待识别文本;

实体识别单元,用于将所述待识别文本中的每一分词输入至实体类型识别模型,得到所述实体类型识别模型输出的每一分词对应的实体类型识别结果;

其中,所述实体类型识别模型是基于同义词交互注意力机制构建,并通过样本文本中的每一样本分词及其样本实体类型标识,以及同义词集字典训练得到的;

所述实体类型识别模型包括输入层、同义词交互注意力层和分类输出层;

对应地,所述实体识别单元具体用于:

将所述待识别文本中的每一分词输入至所述输入层,得到所述输入层输出的每一分词的词向量;

将每一分词的词向量输入至所述同义词交互注意力层,得到所述同义词交互注意力层输出的每一分词的增强词向量;

将每一分词的增强词向量输入至所述分类输出层,得到所述分类输出层输出的每一分词的实体类型识别结果;

其中，所述同义词交互注意力层基于同义词集字典，判断输入的每一分词中是否存在同义词，从而对同义词的词向量进行交互增强，并输出每一分词的增强词向量。

10.一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现如权利要求1至8中任一项所述的文本实体类型识别方法的步骤。

11.一种非暂态计算机可读存储介质，其上存储有计算机程序，其特征在于，该计算机程序被处理器执行时实现如权利要求1至8中任一项所述的文本实体类型识别方法的步骤。

文本实体类型识别方法、装置、电子设备和存储介质

技术领域

[0001] 本发明涉及自然语言处理技术领域，尤其涉及一种文本实体类型识别方法、装置、电子设备和存储介质。

背景技术

[0002] 如今，人工智能和大数据技术在教育领域扮演着重要的角色。在人工智能教育过程中，会产生大量的用于学生考试和练习的试题文本数据，学生作答后会产生更多的作答文本数据，数据量巨大，信息结构复杂。对上述文本数据进行文本实体类型识别是后续智能批改作业、题目难度预测以及题目知识点预测等相关应用的前提。

[0003] 现有的文本实体类型识别，通常是针对中英文文本数据中的人名、地名、机构名或其他具有特定意义的词汇进行实体类型识别，需要识别的实体类型的表达方式一般较为固定。然而试题文本数据，尤其是数学文本数据中，同一类型的文本实体可能存在多种表达方式，文本实体表达方式的多变为文本实体类型识别带来了困难。

发明内容

[0004] 本发明实施例提供一种文本实体类型识别方法、装置、电子设备和存储介质，用以解决由于文本实体多变的表达方式导致文本实体类型识别准确性低的问题。

[0005] 第一方面，本发明实施例提供一种文本实体类型识别方法，包括：

[0006] 确定待识别文本；

[0007] 将所述待识别文本中的每一分词输入至实体类型识别模型，得到所述实体类型识别模型输出的每一分词对应的实体类型识别结果；

[0008] 其中，所述实体类型识别模型是基于同义词交互注意力机制构建，并通过样本文本中的每一样本分词及其样本实体类型标识，以及同义词集字典训练得到的。

[0009] 优选地，所述实体类型识别模型包括输入层、同义词交互注意力层和分类输出层；

[0010] 对应地，所述将所述待识别文本中的每一分词输入至实体类型识别模型，得到所述实体类型识别模型输出的每一分词对应的实体类型识别结果，具体包括：

[0011] 将所述待识别文本中的每一分词输入至所述输入层，得到所述输入层输出的每一分词的词向量；

[0012] 将每一分词的词向量输入至所述同义词交互注意力层，得到所述同义词交互注意力层输出的每一分词的增强词向量；

[0013] 将每一分词的增强词向量输入至所述分类输出层，得到所述分类输出层输出的每一分词的实体类型识别结果。

[0014] 优选地，所述将每一分词的增强词向量输入至所述分类输出层，得到所述分类输出层输出的每一分词的实体类型识别结果，之前还包括：

[0015] 针对任一分词，将所述任一分词的增强词向量更新为所述任一分词的增强词向量和所述任一分词的字符特征向量的拼接向量。

[0016] 优选地，所述实体类型识别模型还包括公式语义预测层；

[0017] 对应地，所述将每一分词的词向量输入至所述同义词交互注意力层，得到所述同义词交互注意力层输出的每一分词的增强词向量，之后还包括：

[0018] 基于每一分词的增强词向量，确定所述待识别文本中公式的词向量；

[0019] 将所述公式的词向量输入至所述公式语义预测层，得到所述公式语义预测层输出的公式语义；

[0020] 对应地，所述实体类型识别模型是基于样本文本中的每一样本分词及其样本实体类型标识，同义词集字典，以及所述样本文本中样本公式的样本公式语义训练得到的。

[0021] 优选地，所述将每一分词的词向量输入至所述同义词交互注意力层，得到所述同义词交互注意力层输出的每一分词的增强词向量，具体包括：

[0022] 基于所述同义词集字典，从每一分词的词向量中选取任一分词的同义词的词向量，并构建所述任一分词的同义词集；所述同义词集包括每一同义词的词向量；

[0023] 基于所述任一分词的词向量与所述同义词集中任一同义词的词向量，确定所述任一分词与所述任一同义词的相似度；

[0024] 基于所述任一分词分别与所述同义词集中每一同义词的相似度，输出所述任一分词的增强词向量。

[0025] 优选地，所述基于所述任一分词分别与所述同义词集中每一同义词的相似度，输出所述任一分词的增强词向量，具体包括：

[0026] 基于所述任一分词分别与所述同义词集中每一同义词的相似度，确定每一同义词的词向量对应的权重；

[0027] 基于所述每一同义词的词向量及其对应的权重，确定注意力词向量；

[0028] 基于所述任一分词的词向量和所述注意力词向量，确定所述任一分词的增强词向量。

[0029] 优选地，所述输出分类层包括上下文层和分类层；

[0030] 对应地，所述将每一分词的增强词向量输入至所述分类输出层，得到所述分类输出层输出的每一分词的实体类型识别结果，具体包括：

[0031] 将每一分词的增强词向量输入至所述上下文层，得到所述上下文层输出的每一分词的序列向量；

[0032] 将每一分词的序列向量输入至所述分类层，得到所述分类层输出的每一分词的实体类型识别结果。

[0033] 优选地，所述分类层包括softmax输出层和随机条件场CRF层；

[0034] 对应地，所述将每一分词的序列向量输入至所述分类层，得到所述分类层输出的每一分词的实体类型识别结果，具体包括：

[0035] 将每一分词的序列向量输入至所述softmax输出层，得到所述softmax输出层输出的每一分词的候选识别结果；

[0036] 将每一分词的候选分类结果输入至所述随机条件场CRF层，得到所述随机条件场CRF层输出的每一分词的实体类型识别结果。

[0037] 优选地，所述待识别文本为数学文本。

[0038] 第二方面，本发明实施例提供一种文本实体类型识别装置，包括：

[0039] 文本确定单元，用于确定待识别文本；

[0040] 实体识别单元，用于将所述待识别文本中的每一分词输入至实体类型识别模型，得到所述实体类型识别模型输出的每一分词对应的实体类型识别结果；

[0041] 其中，所述实体类型识别模型是基于同义词交互注意力机制构建，并通过样本文本中的每一样本分词及其样本实体类型标识，以及同义词集字典训练得到的。

[0042] 第三方面，本发明实施例提供一种电子设备，包括处理器、通信接口、存储器和总线，其中，处理器，通信接口，存储器通过总线完成相互间的通信，处理器可以调用存储器中的逻辑指令，以执行如第一方面所提供的方法的步骤。

[0043] 第四方面，本发明实施例提供一种非暂态计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现如第一方面所提供的方法的步骤。

[0044] 本发明实施例提供的一种文本实体类型识别方法、装置、电子设备和存储介质，通过将待识别文本的每一分词输入至基于同义词交互注意力机制构建的实体类型识别模型中，进行实体类型识别，克服了由于文本实体表达方式的多变导致的识别困难的问题，有助于提高文本实体类型识别的准确性和可靠性。

附图说明

[0045] 为了更清楚地说明本发明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一简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0046] 图1为本发明实施例提供的文本实体类型识别方法的流程示意图；

[0047] 图2为本发明实施例提供的实体类型识别模型预测方法的流程示意图；

[0048] 图3为本发明实施例提供的同义词交互注意力机制的实现方法的流程示意图；

[0049] 图4为本发明实施例提供的实体类型分类输出方法的流程示意图；

[0050] 图5为本发明实施例提供的实体类型识别模型的结构示意图；

[0051] 图6为本发明实施例提供的文本实体类型识别装置的结构示意图；

[0052] 图7为本发明实施例提供的电子设备的结构示意图。

具体实施方式

[0053] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0054] 在人工智能教育过程中，会产生大量的试题文本数据，题量巨大，这些试题应用于学生考试和练习，进而产生更多的作答文本数据。人工智能教育要对每个学生因材施教，就需要对学生的所有历史作答文本数据和试题文本数据进行分析，进而得到学生的详细学情。目前，试题文本数据和作答文本数据，主要由老师、学科编辑或学生产生的，信息结构较为复杂。如何对上述文本数据进行处理、分析和挖掘，是后续智能批改作业、题目难度预测以及题目知识点预测等相关应用的前提。

[0055] 以数学学科为例,数据学科的试题文本数据和作答文本数据中,同一类型的文本实体可能存在多种表达方式,其中一条线段,可以用单个小写字母表示,例如线段a,也可以用线段两个端点对应的大写字母表示,例如线段AB;一个角可以用三个大写字母表示,例如∠AOC,其中O为顶点,还可以用一个大写字母表示,例如∠O,还可以用数字或者希腊字母表示,例如∠β。文本实体表达方式的多变为文本实体类型识别带来了困难。

[0056] 针对上述问题,本发明实施例提供了一种文本实体类型识别方法,该方法既可以用于数学学科的试题文本数据和作答文本数据的文本实体类型识别,也可以用于其他学科,例如物理学科的试题文本数据和作答文本数据的文本实体类型识别,还可以用于论文整理分类等场景。

[0057] 图1为本发明实施例提供的文本实体类型识别方法的流程示意图,如图1所示,该方法包括:

[0058] 步骤110,确定待识别文本。

[0059] 此处,待识别文本即需要进行文本实体类型识别的文本,待识别文本可以是人工录入的,也可以是通过OCR(Optical Character Recognition,光学字符识别)技术对包含有待识别文本的图片进行识别得到的,本发明实施例对此不作具体限定。

[0060] 步骤120,将待识别文本中的每一分词输入至实体类型识别模型,得到实体类型识别模型输出的每一分词对应的实体类型识别结果;其中,实体类型识别模型是基于同义词交互注意力机制构建,并通过样本文本中的每一样本分词及其样本实体类型标识,以及同义词集字典训练得到的。

[0061] 具体地,深度学习中,注意力机制(Attention Mechanism)和人类视觉的注意力机制类似,就是在众多信息中把注意力集中放在重要的点上,选出关键信息,而忽略其他不重要的信息。本发明实施例中,针对文本实体表达方式多变的特性,将同一类型的文本实体的不同表达方式定义为同义词,建立同义词交互注意力机制。同义词交互注意力机制下,针对任一分词,将注意力集中该分词及其同义词的相同信息上,弱化该分词及其同义词之间的不同信息,从而凸显该分词对应文本实体的特征信息,克服了由于文本实体表达方式的多变导致的识别困难。

[0062] 实体类型识别模型用于针对输入的待识别文本中的每一分词,在同义词交互注意力机制下预测每一分词的文本实体类型,并输出每一分词的实体类型识别结果。此处,任一分词的实体类型识别结果即该分词对应的实体类型,或者是该分词对应于每一实体类型的概率。

[0063] 另外,在执行步骤120之前,还可以预先训练得到实体类型识别模型,具体可通过如下方式训练得到实体类型识别模型:

[0064] 首先,收集大量样本文本,对样本文本中的每一样本分词标记对应的样本实体类型标识。此处,样本实体类型标识用于指示样本分词的实体类型。同时,构建同义词集字典。此处,同义词集字典表征同一实体类型的不同表达方式,例如f(x)和g(x)的实体类别均为“一般函数”,故构建“一般函数”的同义词集合E={f(x),g(x)},同义词集字典中包含有大量实体类型的同义词集合。同义词集字典可以是在工作人员阅读、理解样本文本后构建的,也可以是在样本文本中每一样本分词的样本实体类型标识的基础上,经过人工筛选过滤得到的。

[0065] 随即，基于样本文本中的每一样本分词及其样本实体类型标识，以及同义词集字典对初始模型进行训练，从而得到实体类型识别模型。其中，初始模型可以是单一神经网络模型，也可以是多个神经网络模型的组合，本发明实施例不对初始模型的类型和结构作具体限定。

[0066] 本发明实施例提供的方法，通过将待识别文本的每一分词输入至基于同义词交互注意力机制构建的实体类型识别模型中，进行实体类型识别，克服了由于文本实体表达方式的多变导致的识别困难的问题，有助于提高文本实体类型识别的准确性和可靠性。

[0067] 基于上述任一实施例，图2为本发明实施例提供的实体类型识别模型预测方法的流程示意图，如图2所示，该方法中，实体类型识别模型包括输入层、同义词交互注意力层和分类输出层。对应地，步骤120，具体包括：

[0068] 步骤121，将待识别文本中的每一分词输入至输入层，得到输入层输出的每一分词的词向量。

[0069] 具体地，输入层为预先训练得到的语言模型，用于确定分词对应的词向量表示。此处，语言模型可以是长短记忆网络LSTM或者递归神经网络RNN，再或者其他类型的网络，例如ELMO模型。ELMO模型是allenai实验室在2018年提出的，ELMO模型利用语言模型的不同层对分词的不同类型的信息进行编码，连接所有层可以自由组合各种文字表示，可以抓取语义和语法的复杂特征以及能够准确地对不同的上下文进行建模，最终得到较优的词向量。

[0070] 步骤122，将每一分词的词向量输入至同义词交互注意力层，得到同义词交互注意力层输出的每一分词的增强词向量。

[0071] 此处，同义词交互注意力层是基于同义词交互注意力机制构建的，同义词交互注意力层基于同义词集字典，判断输入的每一分词中是否存在同义词，从而对同义词的词向量进行交互增强，以突出同义词之间相同的向量特征，并输出每一分词的增强词向量。此处，增强词向量即经过同义词交互注意力层增强后的词向量。

[0072] 步骤123，将每一分词的增强词向量输入至分类输出层，得到分类输出层输出的每一分词的实体类型识别结果。

[0073] 此处，分类输出层用于对输入的每一分词的增强词向量进行分析预测，判断每一分词对应于实体类型的概率，并输出每一分词的实体类型识别结果。

[0074] 在实际应用中，待识别文本可能存在语言稀疏性的问题。以数学学科为例，数据学科的试题文本数据和作答文本数据中，用于标识数学对象的数学实体极为稀疏，针对这一问题，基于上述任一实施例，该方法中，步骤122和步骤123之间还包括：针对任一分词，将该分词的增强词向量更新为该分词的增强词向量和该分词的字符特征向量的拼接向量。

[0075] 此处，任一分词的字符特征向量用于表征该分词的字符类型，例如在数学文本中，字符类型包括中文、英文、数字和符号四类。且一般来说，中文大多不包含具体的数学对象，英文多作为数学对象的指代，多数的英文均可视为数学实体，而英文与数字、符号经常混合存在，例如常见的数学公式中通常混合有大量的数字、符号以及大小写英文，这导致了实体类型识别难度的增大。对此，本发明实施例对每一分词的字符类型进行编码，生成对应的字符特征向量，并将分词的增强词向量与字符特征向量进行拼接，将拼接后的向量输入分类输出层进行实体类型识别。

[0076] 本发明实施例提供的方法，通过将增强词向量与字符特征向量相结合，进行实体

类型识别，克服了语言稀疏性的问题，能够进一步提高实体类型识别精度。

[0077] 基于上述任一实施例，该方法中，实体类型识别模型还包括公式语义预测层；对应地，步骤122之后还包括：基于每一分词的增强词向量，确定待识别文本中公式的词向量；将公式的词向量输入至公式语义预测层，得到公式语义预测层输出的公式语义。

[0078] 具体地，公式语义是指公式对应的具体数学概率、定理或者被广泛接收的通用数学名词，如公式“f(x)＝x+1”对应的公式语义为“一次函数”，公式“sin(x+y)＝sinx*cosy+cosx*siny”对应的公式语义为“两角和与差的三角函数公式”。

[0079] 同义词交互注意力层输出的每一分词的增强词向量还可以用于待识别文本中公式语义类型的预测。任一公式可能对应一个或多个分词，在已知公式与分词间的对应关系后，可以确定公式对应分词的增强词向量，进而通过求平均或者拼接等方式确定公式的词向量，并将公式的词向量输入到公式语义预测层，由公式语义预测层对公式语义进行预测。

[0080] 本发明实施例中，公式语义的预测和公式中具体的每个分词的实体类型识别之间存在相辅相成的关系，例如当前公式的公式语义为三角函数公式，则当前公式中各个分词的实体类型极可能是三角函数或者角度。在确定公式语义后，公式中对应的分词的实体类型也会被限制在符合公式语义表达的有限空间内，因此，可以基于样本文本中的每一样本分词及其样本实体类型标识，同义词集字典，样本文本中样本公式的样本公式语义训练得到的实体类型识别模型。

[0081] 具体在实体类型识别模型的训练过程中，可以设置两个损失函数，分别用于衡量预测得到的实体类型结果与样本实体类型标识之间的误差，以及预测得到的公式语义和样本公式语义之间的误差，并将两个损失函数的结果按照权值相加，得到实体类型识别模型整体的训练损失。需要说明的是，本发明实施例不对损失函数的权值做具体限定，损失函数的权值大小可以根据对应的任务进行调整，作为优选，在本发明实施例提供的实体类型识别模型中，公式语义预测为辅助任务，公式语义对应的损失函数的权值较小。

[0082] 本发明实施例提供的方法，利用公式语义辅助预测实体类别，可以充分挖掘分词的语境含义，提高实体类型识别的准确性。

[0083] 基于上述任一实施例，图3为本发明实施例提供的同义词交互注意力机制的实现方法的流程示意图，如图3所示，该方法中，步骤122具体包括：

[0084] 步骤1221，基于同义词集字典，从每一分词的词向量中选取任一分词的同义词的词向量，并构建该分词的同义词集；同义词集包括每一同义词的词向量。

[0085] 假设输入至同义词交互注意力层的分词的词向量的数量为n，n个分词分别表示为分词1、分词2、…、分词n，n个分词的词向量分别为词向量1、词向量2、…、词向量n。i为小于等于n的正整数，针对分词i，在预先设定的同义词集字典中查询分词1、分词2、……、分词i-1、分词i+1、……、分词n是否为分词i的同义词，如果是，则将该分词的词向量加入分词i的同义词集中。假设分词2的同义词为分词5和分词10，则分词2的同义词集E*＝{词向量5，词向量10}；假设分词3在当前输入的每一分词中不存在同义词，则分词3的同义词集E*为空集。

[0086] 步骤1222，基于该分词的词向量与同义词集中任一同义词的词向量，确定该分词与该同义词的相似度。

[0087] 假设分词2的同义词集E*＝{词向量5，词向量10}，则分别计算词向量2与词向量5

的相似度，以及词向量2与词向量10的相似度。此处，词向量间相似度的计算可以通过相似度度量公式实现，相似度度量公式包括点积、拼接、或感知机等，本发明实施例对此不作具体限定。

[0088] 步骤1223，基于该分词分别与同义词集中每一同义词的相似度，输出该分词的增强词向量。

[0089] 具体地，可以根据该分词分别与同义词集中每一同义词的相似度，对该分词的词向量与同义词集中同义词的词向量进行融合，从而增强该分词的词向量与同义词的词向量中相同信息，弱化该分词的词向量及其同义词的词向量之间的不同信息，实现对该分词的词向量的同义词交互注意力增强，得到该分词的增强词向量。

[0090] 需要说明的是，若步骤1221中，针对任一分词的同义词集为空集，即待识别文本中不包含该分词的同义词，则同义词交互注意力层输出的该分词的增强词向量仍为该分词的词向量。

[0091] 基于上述任一实施例，该方法中，步骤1223具体包括：基于该分词分别与同义词集中每一同义词的相似度，确定每一同义词的词向量对应的权重；基于每一同义词的词向量及其对应的权重，确定注意力词向量；基于该分词的词向量和注意力词向量，确定该分词的增强词向量。

[0092] 具体地，针对任一分词，将该分词与同义词集中每一同义词的相似度进行权重归一化，得到每一同义词的词向量对应的权重。随后，对每一同义词的词向量进行加权求和，将加权求和的结果作为注意力词向量。最后，该分词的词向量和注意力词向量进行融合，得到该分词的增量词向量，此处，融合的方式可以是求平均或者拼接等。

[0093] 例如，分词2的同义词集E*＝{词向量5，词向量10}，则词向量2与词向量5的相似度为75％，词向量2与词向量10的相似度为85％，将两个相似度进行权重归一化，得到词向量5的权重＝75％/(75％+85％)＝0.46875，词向量10的权重＝85％/(75％+85％)＝0.53125。由此，得到分词2的注意力词向量为0.46875*词向量5+0.53125*词向量10。在此基础上，对词向量2和注意力词向量进行平均，得到分词2的增强词向量。

[0094] 基于上述任一实施例，图4为本发明实施例提供的实体类型分类输出方法的流程示意图，如图4所示，该方法中，输出分类层包括上下文层和分类层；对应地，步骤123具体包括：

[0095] 步骤1231，将每一分词的增强词向量输入至上下文层，得到上下文层输出的每一分词的序列向量。

[0096] 步骤1232，将每一分词的序列向量输入至分类层，得到分类层输出的每一分词的实体类型识别结果。

[0097] 具体地，上下文层用于根据输入的每一分词的增强词向量，分析每一分词的上下文信息，并输出每一分词的序列向量。分类层用于对输入的每一分词的序列向量进行分析预测，判断每一分词对应于实体类型的概率，并输出每一分词的实体类型识别结果。

[0098] 此处，序列向量既包含有分词本身的信息，也包含有分词的上下文信息。上下文层可以通过长短记忆网络LSTM实现，作为优选，本发明实施例中的上下文层为两层双向长短记忆网络Bi-LSTM。

[0099] 基于上述任一实施例，该方法中，分类层包括softmax输出层和随机条件场CRF层；

对应地，步骤1232具体包括：将每一分词的序列向量输入至softmax输出层，得到softmax输出层输出的每一分词的候选识别结果；将每一分词的候选分类结果输入至随机条件场CRF层，得到随机条件场CRF层输出的每一分词的实体类型识别结果。

[0100] 具体地，softmax输出层基于输入的分词的序列向量，确定分词对应于每一实体类型的概率，并将分词对应于每一实体类型的概率作为该分词的候选识别结果输出。随机条件场CRF层基于任一分词的上一分词的候选识别结果，对该分词的候选识别结果进行限制调整，并将调整后的候选识别结果作为该分词的实体类型识别结果输出。

[0101] 由于一个实体可能是由一个或多个分词组成的，本发明实施例采用BIO(Begin，inside，outside)标注形式进行识别结果的标注，其中B表示分词处于一个实体的开始(Begin)，I表示分词处于实体内部(inside)，O表示分词处于实体外部(outside)，即不属于该实体。在随机条件场CRF层中，假设任一分词的候选识别结果为三角形的开始“B-angle”，则该分词的下一分词不可能为圆这一类型的实体“I-circle”。随机条件场CRF层可以通过学习实体间的约束关系，减少无效预测实体序列的数量。

[0102] 基于上述任一实施例，该方法中，待识别文本为数学文本。此处，数学文本是完整的数学题目信息，包括题干、答案和解析三个部分。

[0103] 图5为本发明实施例提供的实体类型识别模型的结构示意图，参考图5，当待识别文本为数学文本时，文本实体类型识别方法具体包括：

[0104] 首先，对待识别的数学文本进行预处理，此处的预处理主要指分词处理。分词可以采用基于规则的方法、基于序列标注的模型预测方法等。由于数学文本包含中文，英文和数字以及符号，本发明实施例将基于规则的方法和中文jieba分词工具进行融合得到数学文本的分词结果。

[0105] 完成分词后，以数学文本的分句为单位，将任一分句中的每一分词，以及每一分词的用于表征分词字符类型的字符特征向量输入到实体类型识别模型中，以获取该分句中实体类型识别结果。

[0106] 此处，实体类型识别模型包括输入层、同义词交互注意力层、上下文层、Softmax输出层、随机条件场CRF层以及公式语义预测层。其中，输入层为预先训练得到的语言模型，用于确定分词对应的词向量；同义词交互注意力层基于同义词集字典，判断输入的每一分词在该分句中是否存在同义词，从而对同义词的词向量进行交互增强，并输出每一分词的增强词向量；上下文层为两层双向长短记忆网络Bi-LSTM，用于从每一分词的增强词向量和字符特征向量的拼接向量中，提取每一分词的上下文信息，并输出每一分词的序列向量；softmax输出层基于输入的分词的序列向量，确定分词对应于每一实体类型的概率，并将分词对应于每一实体类型的概率作为该分词的候选识别结果输出。随机条件场CRF层对每一分词的候选识别结果进行限制调整，并将调整后的候选识别结果作为该分词的实体类型识别结果输出；公式语义预测层用于根据该分句中公式的词向量，预测并输出公式语义。此处，公式的词向量是根据公式所对应的分词的增强词向量确定的。

[0107] 本发明实施例中，利用同义词交互注意力机制来增强词向量，将增强词向量和数学文本独有的字符特征向量相结合，并且利用公式语义辅助预测实体类别，可以充分挖掘分词的语境含义和分词特征，既能较好的解决数学文本的语言稀疏性问题，也能较好的解决实体类别数量不均衡情况下的实体识别问题。

[0108] 基于上述任一实施例，本发明实施例提供一种实体类型识别模型的训练方法：

[0109] 首先，收集大量样本数学文本，对样本数学文本中的每一样本分词标记对应的样本实体类型标识。针对于数学实体，其类型数量远远超过了常见的中英文命名实体识别任务，故本发明实施例按照层级关系对数学实体类型进行定义，如定义数学实体“四边形”，“四边形”具体可以包含“长方形”，“正方形”以及“平行四边形”等数学实体。在执行实体识别时，可以根据具体需求将数学实体标签进行归一化，即可得到不同粒度的数学实体类型识别结果。

[0110] 同时，构建同义词集字典。同义词集字典可以是在工作人员阅读、理解样本数据文本后构建的，也可以是在样本数据文本中每一样本分词的样本实体类型标识的基础上，经过人工筛选，过滤非同义词后得到的。

[0111] 此外，还需要标注样本数学文本中公式部分对应的公式语义。

[0112] 随即，构建一种由输入层、同义词交互注意力层、上下文层、Softmax输出层、随机条件场CRF层以及公式语义预测层构成的神经网络结构，将数学实体类型识别作为主要任务，将预测文本中公式的语义作为辅助任务，采用多任务学习的方式训练得到样本数学文本中每一分词对应于每一数学实体类型的概率分布。

[0113] 基于上述任一实施例，图6为本发明实施例提供的文本实体类型识别装置的结构示意图，如图6所示，该装置包括文本确定单元610和实体识别单元620；

[0114] 其中，文本确定单元610用于确定待识别文本；

[0115] 实体识别单元620用于将所述待识别文本中的每一分词输入至实体类型识别模型，得到所述实体类型识别模型输出的每一分词对应的实体类型识别结果；

[0116] 其中，所述实体类型识别模型是基于同义词交互注意力机制构建，并通过样本文本中的每一样本分词及其样本实体类型标识，以及同义词集字典训练得到的。

[0117] 本发明实施例提供的装置，通过将待识别文本的每一分词输入至基于同义词交互注意力机制构建的实体类型识别模型中，进行实体类型识别，克服了由于文本实体表达方式的多变导致的识别困难的问题，有助于提高文本实体类型识别的准确性和可靠性。

[0118] 基于上述任一实施例，该装置中，所述实体类型识别模型包括输入层、同义词交互注意力层和分类输出层；

[0119] 对应地，所述实体识别单元620包括：

[0120] 输入子单元，用于将所述待识别文本中的每一分词输入至所述输入层，得到所述输入层输出的每一分词的词向量；

[0121] 注意力子单元，用于将每一分词的词向量输入至所述同义词交互注意力层，得到所述同义词交互注意力层输出的每一分词的增强词向量；

[0122] 分类输出子单元，用于将每一分词的增强词向量输入至所述分类输出层，得到所述分类输出层输出的每一分词的实体类型识别结果。

[0123] 基于上述任一实施例，该装置中，所述实体识别单元620还包括：

[0124] 向量拼接子单元，用于针对任一分词，将所述任一分词的增强词向量更新为所述任一分词的增强词向量和所述任一分词的字符特征向量的拼接向量。

[0125] 基于上述任一实施例，该装置中，所述实体类型识别模型还包括公式语义预测层；

[0126] 对应地，所述实体识别单元620还包括：

[0127] 语义预测子单元，用于基于每一分词的增强词向量，确定所述待识别文本中公式的词向量；

[0128] 将所述公式的词向量输入至所述公式语义预测层，得到所述公式语义预测层输出的公式语义；

[0129] 对应地，所述实体类型识别模型是基于样本文本中的每一样本分词及其样本实体类型标识，同义词集字典，以及所述样本文本中样本公式的样本公式语义训练得到的。

[0130] 基于上述任一实施例，该装置中，所述注意力子单元包括：

[0131] 同义词确定模块，用于基于所述同义词集字典，从每一分词的词向量中选取任一分词的同义词的词向量，并构建所述任一分词的同义词集；所述同义词集包括每一同义词的词向量；

[0132] 相似度确定模块，用于基于所述任一分词的词向量与所述同义词集中任一同义词的词向量，确定所述任一分词与所述任一同义词的相似度；

[0133] 向量增强模块，用于基于所述任一分词分别与所述同义词集中每一同义词的相似度，输出所述任一分词的增强词向量。

[0134] 基于上述任一实施例，该装置中，所述向量增强模块具体用于：

[0135] 基于所述任一分词分别与所述同义词集中每一同义词的相似度，确定每一同义词的词向量对应的权重；

[0136] 基于所述每一同义词的词向量及其对应的权重，确定注意力词向量；

[0137] 基于所述任一分词的词向量和所述注意力词向量，确定所述任一分词的增强词向量。

[0138] 基于上述任一实施例，该装置中，所述输出分类层包括上下文层和分类层；

[0139] 对应地，所述分类输出子单元包括：

[0140] 上下文模块，用于将每一分词的增强词向量输入至所述上下文层，得到所述上下文层输出的每一分词的序列向量；

[0141] 分类模块，用于将每一分词的序列向量输入至所述分类层，得到所述分类层输出的每一分词的实体类型识别结果。

[0142] 基于上述任一实施例，该装置中，所述分类层包括softmax输出层和随机条件场CRF层；

[0143] 对应地，所述分类模块具体用于：

[0144] 将每一分词的序列向量输入至所述softmax输出层，得到所述softmax输出层输出的每一分词的候选识别结果；

[0145] 将每一分词的候选分类结果输入至所述随机条件场CRF层，得到所述随机条件场CRF层输出的每一分词的实体类型识别结果。

[0146] 基于上述任一实施例，该装置中，所述待识别文本为数学文本。

[0147] 图7为本发明实施例提供的电子设备的结构示意图，如图7所示，该电子设备可以包括：处理器(processor)710、通信接口(Communications Interface)720、存储器(memory)730和通信总线740，其中，处理器710，通信接口720，存储器730通过通信总线740完成相互间的通信。处理器710可以调用存储器730中的逻辑指令，以执行如下方法：确定待识别文本；将所述待识别文本中的每一分词输入至实体类型识别模型，得到所述实体类型

识别模型输出的每一分词对应的实体类型识别结果;其中,所述实体类型识别模型是基于同义词交互注意力机制构建,并通过样本文本中的每一样本分词及其样本实体类型标识,以及同义词集字典训练得到的。

[0148] 此外,上述的存储器730中的逻辑指令可以通过软件功能单元的形式实现并作为独立的产品销售或使用时,可以存储在一个计算机可读取存储介质中。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0149] 本发明实施例还提供一种非暂态计算机可读存储介质,其上存储有计算机程序,该计算机程序被处理器执行时实现以执行上述各实施例提供的方法,例如包括:确定待识别文本;将所述待识别文本中的每一分词输入至实体类型识别模型,得到所述实体类型识别模型输出的每一分词对应的实体类型识别结果;其中,所述实体类型识别模型是基于同义词交互注意力机制构建,并通过样本文本中的每一样本分词及其样本实体类型标识,以及同义词集字典训练得到的。

[0150] 以上所描述的装置实施例仅仅是示意性的,其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下,即可以理解并实施。

[0151] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到各实施方式可借助软件加必需的通用硬件平台的方式来实现,当然也可以通过硬件。基于这样的理解,上述技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品可以存储在计算机可读存储介质中,如ROM/RAM、磁碟、光盘等,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行各个实施例或者实施例的某些部分所述的方法。

[0152] 最后应说明的是:以上实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

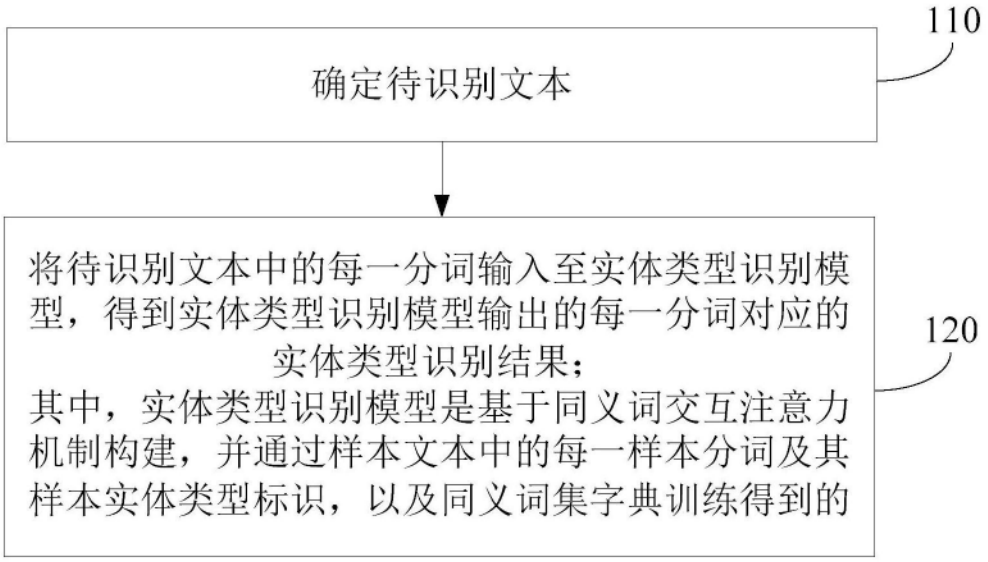

110
确定待识别文本
120
将待识别文本中的每一分词输入至实体类型识别模型，得到实体类型识别模型输出的每一分词对应的实体类型识别结果；
其中，实体类型识别模型是基于同义词交互注意力机制构建，并通过样本文本中的每一样本分词及其样本实体类型标识，以及同义词集字典训练得到的

图1

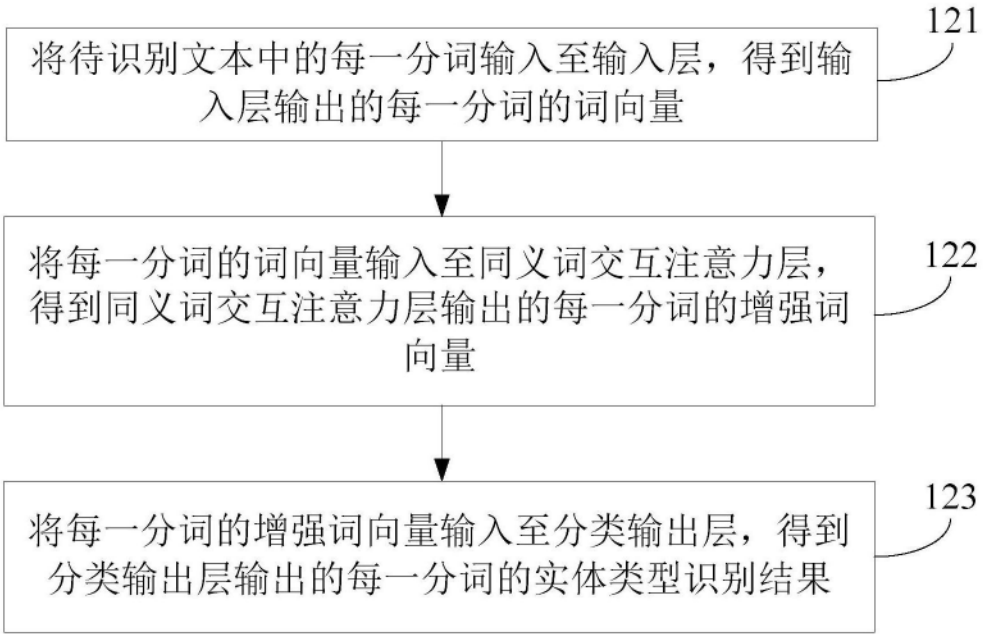

121
将待识别文本中的每一分词输入至输入层，得到输入层输出的每一分词的词向量
122
将每一分词的词向量输入至同义词交互注意力层，得到同义词交互注意力层输出的每一分词的增强词向量
123
将每一分词的增强词向量输入至分类输出层，得到分类输出层输出的每一分词的实体类型识别结果

图2

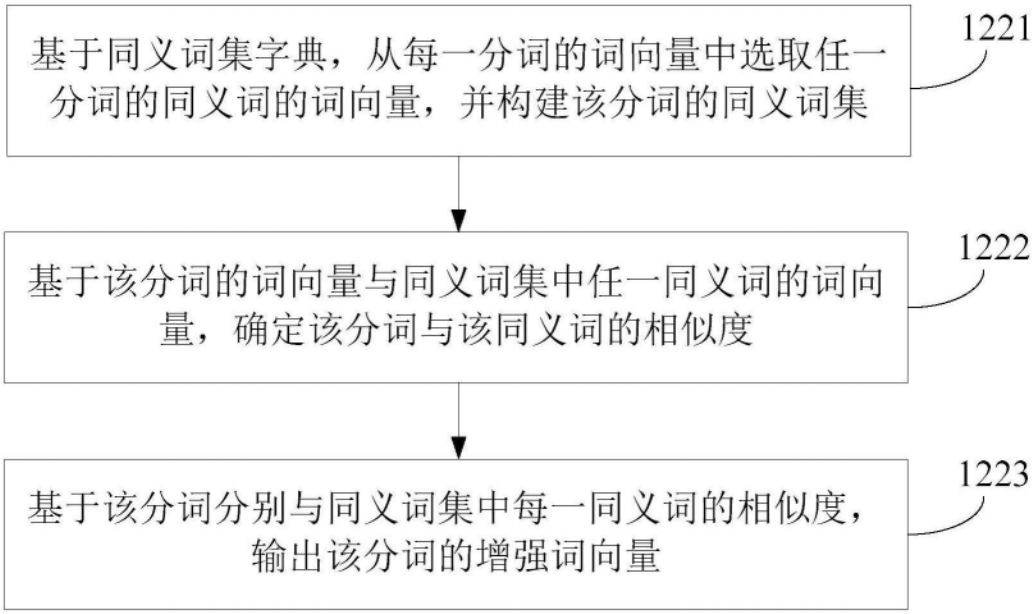

基于同义词集字典，从每一分词的词向量中选取任一分词的同义词的词向量，并构建该分词的同义词集
1221
基于该分词的词向量与同义词集中任一同义词的词向量，确定该分词与该同义词的相似度
1222
基于该分词分别与同义词集中每一同义词的相似度，输出该分词的增强词向量
1223

图3

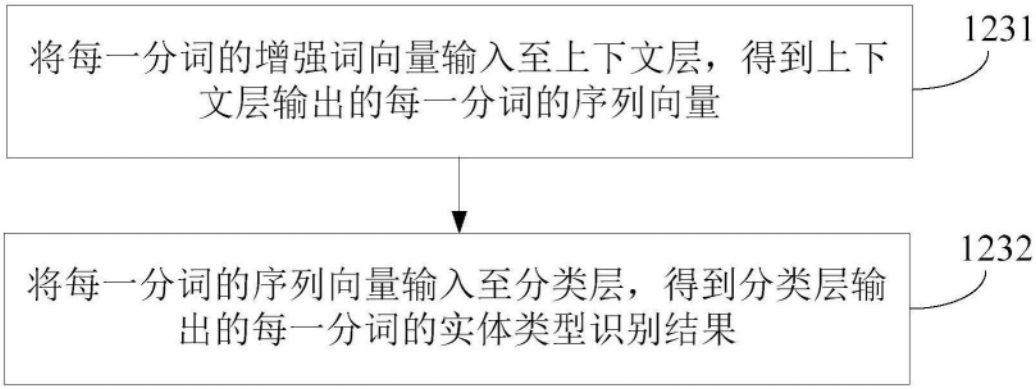

将每一分词的增强词向量输入至上下文层，得到上下文层输出的每一分词的序列向量
1231
将每一分词的序列向量输入至分类层，得到分类层输出的每一分词的实体类型识别结果
1232

图4

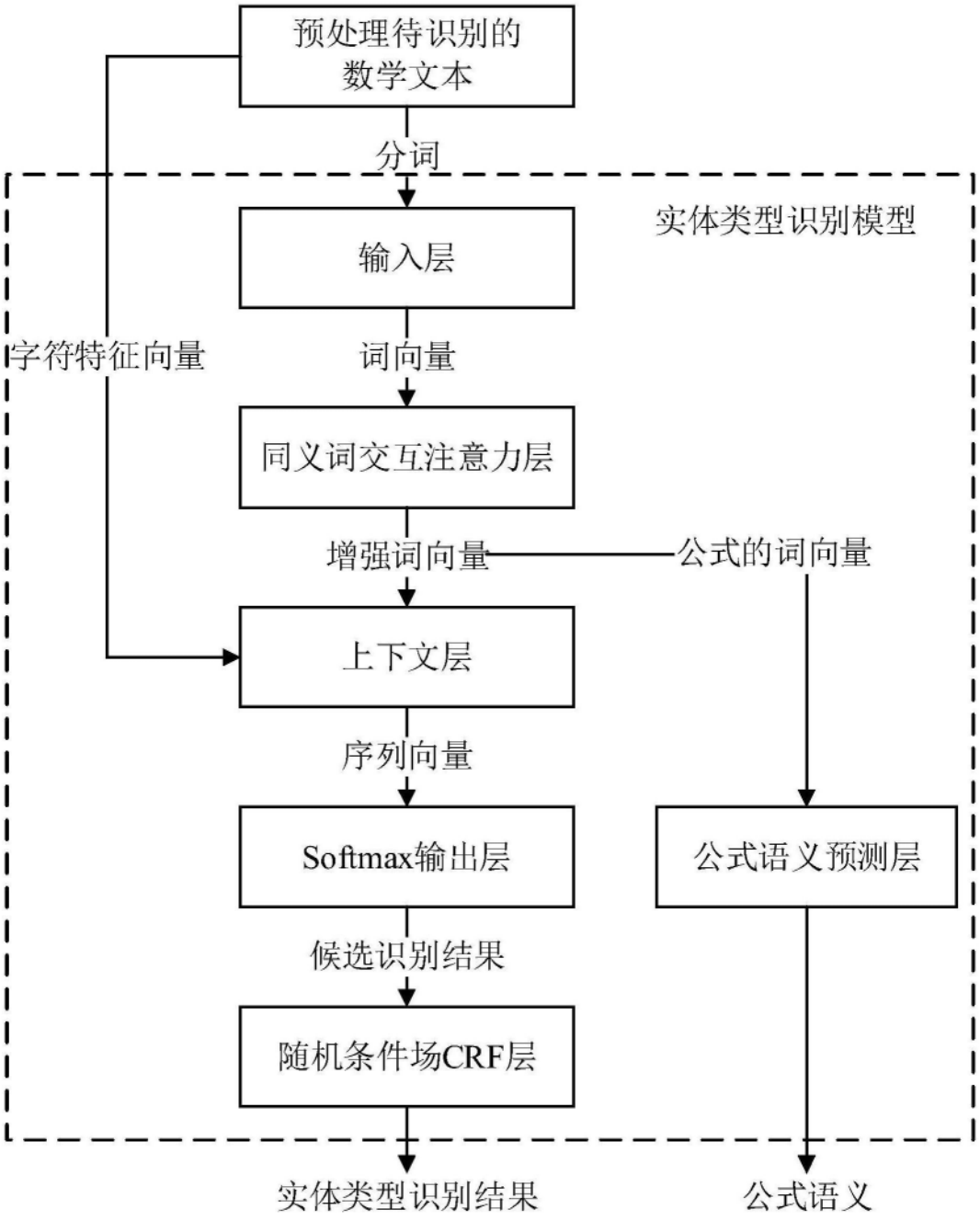

预处理待识别的
数学文本
分词
实体类型识别模型
输入层
字符特征向量
词向量
同义词交互注意力层
增强词向量
公式的词向量
上下文层
序列向量
Softmax输出层
公式语义预测层
候选识别结果
随机条件场CRF层
实体类型识别结果
公式语义

图5

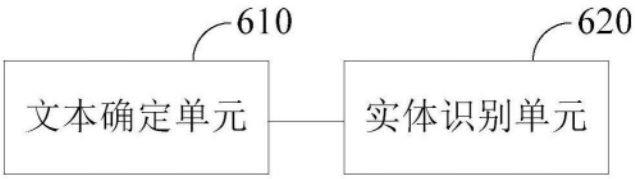

610
620
文本确定单元
实体识别单元

图6

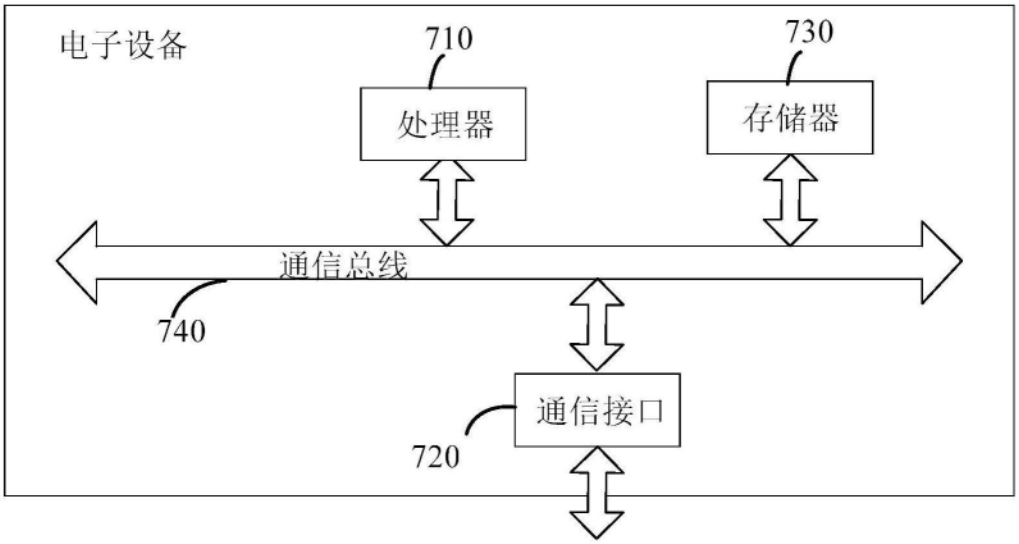


图7