(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111291185 B
(45)授权公告日 2023.09.22

(21)申请号 202010071824.3
(22)申请日 2020.01.21
(65)同一申请的已公布的文献号
申请公布号 CN 111291185 A
(43)申请公布日 2020.06.16
(73)专利权人 京东方科技集团股份有限公司
地址 100015 北京市朝阳区酒仙桥路10号
(72)发明人 王炳乾
(74)专利代理机构 北京润泽恒知识产权代理有限公司 11319
专利代理师 李娜
(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/36(2019.01)
G06N 3/0464(2023.01)
G06N 3/08(2023.01)

(56)对比文件
CN 108874778 A,2018.11.23
CN 106844368 A,2017.06.13
CN 106709006 A,2017.05.24
CN 108694208 A,2018.10.23
CN 106055536 A,2016.10.26
US 10325106 B1,2019.06.18
US 2019087724 A1,2019.03.21
US 2018082183 A1,2018.03.22

审查员 张俊

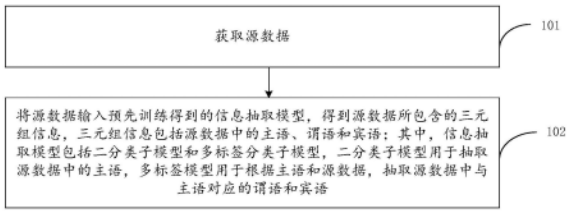
权利要求书3页 说明书10页 附图5页

(54)发明名称

信息抽取方法、装置、电子设备及存储介质

(57)摘要

本申请提供了一种信息抽取方法、装置、电子设备及存储介质，首先获取源数据，然后将源数据输入预先训练得到的信息抽取模型，得到源数据所包含的三元组信息，三元组信息包括源数据中的主语、谓语和宾语；其中，信息抽取模型包括二分类子模型和多标签分类子模型，二分类子模型用于抽取源数据中的主语，多标签分类子模型用于根据主语和所述源数据，抽取源数据中与主语对应的谓语和宾语。本申请技术方案采用端到端的信息抽取模型联合抽取源数据中的三元组信息，替代传统的实体识别和关系抽取的管道式抽取方法，可以提高信息抽取的效率和准确率。



CN 111291185 B

1.一种信息抽取方法,其特征在于,所述方法包括:

获取源数据;

将所述源数据输入预先训练得到的信息抽取模型,得到所述源数据所包含的三元组信息,所述三元组信息包括所述源数据中的主语、谓语和宾语;其中,所述信息抽取模型包括二分类子模型和多标签分类子模型,所述二分类子模型用于抽取所述源数据中的主语,所述多标签分类子模型用于根据所述主语和所述源数据,抽取所述源数据中与所述主语对应的谓语和宾语;所述源数据包括非结构化文本,所述二分类子模型和所述多标签分类子模型是采用样本集合对预训练语言模型和神经网络模型进行联合训练得到,所述样本集合包括多个待训练文本以及各所述待训练文本的三元组标注信息,所述三元组标注信息包括主语标注信息、谓语标注信息和宾语标注信息;

在所述将所述源数据输入预先训练得到的信息抽取模型,得到所述源数据所包含的三元组信息的步骤之前,还包括:获得所述信息抽取模型,其中,所述获得所述信息抽取模型的步骤,包括:

获得所述样本集合;

将所述待训练文本输入第一预训练语言模型,将所述第一预训练语言模型的输出信息送入第一神经网络模型;

将所述第一神经网络模型的输出信息以及所述待训练文本输入第二预训练语言模型,将所述第二预训练语言模型的输出信息送入第二神经网络模型;

根据所述第一神经网络模型的输出信息、所述第二神经网络模型的输出信息以及所述三元组标注信息,对所述第一预训练语言模型、所述第一神经网络模型、所述第二预训练语言模型以及所述第二神经网络模型进行训练,得到所述信息抽取模型,其中,训练后的第一预训练语言模型和第一神经网络模型构成所述二分类子模型,训练后的第二预训练语言模型和第二神经网络模型构成所述多标签分类子模型。

2.根据权利要求1所述的信息抽取方法,其特征在于,所述根据所述第一神经网络模型的输出信息、所述第二神经网络模型的输出信息以及所述三元组标注信息,对所述第一预训练语言模型、所述第一神经网络模型、所述第二预训练语言模型以及所述第二神经网络模型进行训练,得到所述信息抽取模型的步骤,包括:

根据所述第一神经网络模型的输出信息以及所述主语标注信息,确定第一损失函数;

根据所述第二神经网络模型的输出信息、所述谓语标注信息以及所述宾语标注信息,确定第二损失函数;

对所述第一预训练语言模型、所述第一神经网络模型、所述第二预训练语言模型以及所述第二神经网络模型中的参数进行优化,得到所述信息抽取模型,使得所述第一损失函数与所述第二损失函数之和最小。

3.根据权利要求2所述的信息抽取方法,其特征在于,所述第一损失函数和所述第二损失函数均为交叉熵损失函数。

4.根据权利要求1所述的信息抽取方法,其特征在于,所述获得样本集合的步骤,包括:

获取非结构化文本样本;

对所述非结构化文本样本进行处理,得到待标注文本;

获取已完成标注的待训练文本以及所述待训练文本的三元组标注信息;

响应于所述待标注文本中包含所述三元组标注信息中的主语标注信息和宾语标注信息，按照所述三元组标注信息对所述待标注文本进行标注。

5.根据权利要求4所述的信息抽取方法，其特征在于，所述获得样本集合的步骤，还包括：

采用预先训练得到的K个预测模型对所述待标注文本进行预测，得到K个三元组预测信息；

当第一三元组信息的数量与K的比值大于第一预设阈值时，将所述第一三元组信息作为所述待标注文本的三元组标注信息添加至所述样本集合中，其中，所述第一三元组信息为出现在所述三元组预测信息中但未出现在所述待标注文本的三元组标注信息中的三元组信息；

当第二三元组信息的数量与K的比值大于第二预设阈值时，将所述第二三元组信息从所述待标注文本的三元组标注信息中删除，其中，所述第二三元组信息为出现在所述待标注文本的三元组标注信息中但未出现在所述三元组预测信息中的三元组信息；

其中，K大于或等于5且小于或等于10。

6.据权利要求5所述的信息抽取方法，其特征在于，在所述采用预先训练得到的K个预测模型对所述待标注文本进行预测，得到K个三元组预测信息的步骤之前，包括：

根据已完成标注的待训练文本以及所述待训练文本的三元组标注信息，采用K折交叉验证的方式获得K个预测模型。

7.一种信息抽取装置，其特征在于，所述装置包括：

获取模块，被配置为获取源数据；

抽取模块，被配置为将所述源数据输入预先训练得到的信息抽取模型，得到所述源数据所包含的三元组信息，所述三元组信息包括所述源数据中的主语、谓语和宾语；其中，所述信息抽取模型包括二分类子模型和多标签分类子模型，所述二分类子模型用于抽取所述源数据中的主语，所述多标签分类子模型用于根据所述主语和所述源数据，抽取所述源数据中与所述主语对应的谓语和宾语；所述源数据包括非结构化文本，所述二分类子模型和所述多标签分类子模型是采用样本集合对预训练语言模型和神经网络模型进行联合训练得到，所述样本集合包括多个待训练文本以及各所述待训练文本的三元组标注信息，所述三元组标注信息包括主语标注信息、谓语标注信息和宾语标注信息；

所述装置还包括：模型获取模块，被配置为获得所述信息抽取模型，所述模型获取模块包括：

第一单元，被配置为获得样本集合；

第二单元，被配置为将所述待训练文本输入第一预训练语言模型，将所述第一预训练语言模型的输出信息送入第一神经网络模型；

第三单元，被配置为将所述第一神经网络模型的输出信息以及所述待训练文本输入第二预训练语言模型，将所述第二预训练语言模型的输出信息送入第二神经网络模型；

第四单元，被配置为根据所述第一神经网络模型的输出信息、所述第二神经网络模型的输出信息以及所述三元组标注信息，对所述第一预训练语言模型、所述第一神经网络模型、所述第二预训练语言模型以及所述第二神经网络模型进行训练，得到所述信息抽取模型，其中，训练后的第一预训练语言模型和第一神经网络模型构成所述二分类子模型，训练

后的第二预训练语言模型和第二神经网络模型构成所述多标签分类子模型。

8.一种电子设备,其特征在于,所述电子设备包括:

处理器;

用于存储所述处理器可执行指令的存储器;

其中,所述处理器被配置为执行所述指令,以实现如权利要求1至6中任一项所述的信息抽取方法。

9.一种存储介质,其特征在于,当所述存储介质中的指令由电子设备的处理器执行时,使得所述电子设备能够执行如权利要求1至6中任一项所述的信息抽取方法。

信息抽取方法、装置、电子设备及存储介质

技术领域

[0001] 本发明涉及信息处理技术领域，特别是涉及一种信息抽取方法、装置、电子设备及存储介质。

背景技术

[0002] 随着深度学习等领域的持续发展，人工智能逐渐涉足各个领域，致力于改善人们的生活，在图像识别、语音识别等领域已经超越了人类的水平。然而在自然语言处理领域，由于人类语言的复杂性以及事物的多样性，目前的技术尚不能达到完全理解语义的程度，因此需要一个语义连接的桥梁——知识图谱。知识图谱由实体、属性和关系组成，其本质上来讲是一种语义网络，网络中的节点表示现实世界存在的实体或者属性值，节点之间的边表示两个实体之间的关系。目前知识图谱技术主要用于智能语义搜索、移动个人助理以及问答系统中。

发明内容

[0003] 本发明提供一种信息抽取方法、装置、电子设备及存储介质，以提高信息抽取的效率和精度。

[0004] 为了解决上述问题，本发明公开了一种信息抽取方法，所述方法包括：

[0005] 获取源数据；

[0006] 将所述源数据输入预先训练得到的信息抽取模型，得到所述源数据所包含的三元组信息，所述三元组信息包括所述源数据中的主语、谓语和宾语；其中，所述信息抽取模型包括二分类子模型和多标签分类子模型，所述二分类子模型用于抽取所述源数据中的主语，所述多标签分类子模型用于根据所述主语和所述源数据，抽取所述源数据中与所述主语对应的谓语和宾语。

[0007] 在一种可选的实现方式中，在所述将所述源数据输入预先训练得到的信息抽取模型，得到所述源数据所包含的三元组信息的步骤之前，还包括：获得所述信息抽取模型，其中，所述获得所述信息抽取模型的步骤，包括：

[0008] 获得样本集合，所述样本集合中包括多个待训练文本以及各所述待训练文本的三元组标注信息，所述三元组标注信息包括主语标注信息、谓语标注信息和宾语标注信息；

[0009] 将所述待训练文本输入第一预训练语言模型，将所述第一预训练语言模型的输出信息送入第一神经网络模型；

[0010] 将所述第一神经网络模型的输出信息以及所述待训练文本输入第二预训练语言模型，将所述第二预训练语言模型的输出信息送入第二神经网络模型；

[0011] 根据所述第一神经网络模型的输出信息、所述第二神经网络模型的输出信息以及所述三元组标注信息，对所述第一预训练语言模型、所述第一神经网络模型、所述第二预训练语言模型以及所述第二神经网络模型进行训练，得到所述信息抽取模型，其中，训练后的第一预训练语言模型和第一神经网络模型构成所述二分类子模型，训练后的第二预训练语

言模型和第二神经网络模型构成所述多标签分类子模型。

[0012] 在一种可选的实现方式中，所述根据所述第一神经网络模型的输出信息、所述第二神经网络模型的输出信息以及所述三元组标注信息，对所述第一预训练语言模型、所述第一神经网络模型、所述第二预训练语言模型以及所述第二神经网络模型进行训练，得到所述信息抽取模型的步骤，包括：

[0013] 根据所述第一神经网络模型的输出信息以及所述主语标注信息，确定第一损失函数；

[0014] 根据所述第二神经网络模型的输出信息、所述谓语标注信息以及所述宾语标注信息，确定第二损失函数；

[0015] 对所述第一预训练语言模型、所述第一神经网络模型、所述第二预训练语言模型以及所述第二神经网络模型中的参数进行优化，得到所述信息抽取模型，使得所述第一损失函数与所述第二损失函数之和最小。

[0016] 在一种可选的实现方式中，所述第一损失函数和所述第二损失函数均为交叉熵损失函数。

[0017] 在一种可选的实现方式中，所述获得样本集合的步骤，包括：

[0018] 获取非结构化文本样本；

[0019] 对所述非结构化文本样本进行处理，得到待标注文本；

[0020] 获取已完成标注的待训练文本以及所述待训练文本的三元组标注信息；

[0021] 响应于所述待标注文本中包含所述三元组标注信息中的主语标注信息和宾语标注信息，按照所述三元组标注信息对所述待标注文本进行标注。

[0022] 在一种可选的实现方式中，所述获得样本集合的步骤，还包括：

[0023] 采用预先训练得到的K个预测模型对所述待标注文本进行预测，得到K个三元组预测信息；

[0024] 当第一三元组信息的数量与K的比值大于第一预设阈值时，将所述第一三元组信息作为所述待标注文本的三元组标注信息添加至所述样本集合中，其中，所述第一三元组信息为出现在所述三元组预测信息中但未出现在所述待标注文本的三元组标注信息中的三元组信息；

[0025] 当第二三元组信息的数量与K的比值大于第二预设阈值时，将所述第二三元组信息从所述待标注文本的三元组标注信息中删除，其中，所述第二三元组信息为出现在所述待标注文本的三元组标注信息中但未出现在所述三元组预测信息中的三元组信息；

[0026] 其中，K大于或等于5且小于或等于10。

[0027] 在一种可选的实现方式中，在所述采用预先训练得到的K个预测模型对所述待标注文本进行预测，得到K个三元组预测信息的步骤之前，包括：

[0028] 根据已完成标注的待训练文本以及所述待训练文本的三元组标注信息，采用K折交叉验证的方式获得K个预测模型。

[0029] 为了解决上述问题，本发明还公开了一种信息抽取装置，所述装置包括：

[0030] 获取模块，被配置为获取源数据；

[0031] 抽取模块，被配置为将所述源数据输入预先训练得到的信息抽取模型，得到所述源数据所包含的三元组信息，所述三元组信息包括所述源数据中的主语、谓语和宾语；其

中，所述信息抽取模型包括二分类子模型和多标签分类子模型，所述二分类子模型用于抽取所述源数据中的主语，所述多标签分类子模型用于根据所述主语和所述源数据，抽取所述源数据中与所述主语对应的谓语和宾语。

[0032] 为了解决上述问题，本发明还公开了一种电子设备，所述电子设备包括：

[0033] 处理器；

[0034] 用于存储所述处理器可执行指令的存储器；

[0035] 其中，所述处理器被配置为执行所述指令，以实现任一实施例所述的信息抽取方法。

[0036] 为了解决上述问题，本发明还公开了一种存储介质，当所述存储介质中的指令由电子设备的处理器执行时，使得所述电子设备能够执行任一实施例所述的信息抽取方法。

[0037] 与现有技术相比，本发明包括以下优点：

[0038] 本申请技术方案提供了一种信息抽取方法、装置、电子设备及存储介质，首先获取源数据，然后将源数据输入预先训练得到的信息抽取模型，得到源数据所包含的三元组信息，三元组信息包括源数据中的主语、谓语和宾语；其中，信息抽取模型包括二分类子模型和多标签分类子模型，二分类子模型用于抽取源数据中的主语，多标签分类子模型用于根据主语和所述源数据，抽取源数据中与主语对应的谓语和宾语。本申请技术方案采用端到端的信息抽取模型联合抽取源数据中的三元组信息，替代传统的实体识别和关系抽取的管道式抽取方法，可以提高信息抽取的效率和准确率。

附图说明

[0039] 为了更清楚地说明本发明实施例的技术方案，下面将对本发明实施例的描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅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性的前提下，还可以根据这些附图获得其他的附图。

[0040] 图1示出了本申请一实施例提供的一种信息抽取方法的步骤流程图；

[0041] 图2示出了本申请一实施例提供的一种获得信息抽取模型的步骤流程图；

[0042] 图3示出了本申请一实施例提供的一种三元组标注信息的格式；

[0043] 图4示出了本申请一实施例提供的一种信息抽取模型的训练框架；

[0044] 图5示出了本申请一实施例提供的一种数据自动标注方法的步骤流程图；

[0045] 图6示出了本申请一实施例提供的一种自动化标注的流程示意图；

[0046] 图7示出了本申请一实施例提供的一种信息抽取方法的流程示意图；

[0047] 图8示出了本申请一实施例提供的一种信息抽取装置的结构框图。

具体实施方式

[0048] 为使本发明的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本发明作进一步详细的说明。

[0049] 领域知识图谱是从特定领域的特定资源中抽取实体和实体之间的关系，从而构建知识库，它包含的知识体系通常具有很强的领域针对性和专业性。领域知识图谱自上而下进行构建，主要包括schema设计，实体识别、关系抽取、实体链接、知识融合、知识计算等几

个环节。其关键是如何自动抽取信息得到候选知识单元，其中涉及的技术包括：实体抽取、关系抽取和属性抽取，统称为信息抽取。信息抽取也称三元组(S,P,O)抽取，其中S和O为句子的主语和宾语，对应知识图谱中的实体或者属性值，P为谓语，对应实体之间的关系。发明人发现，现有方法大多数将信息抽取分成两个步骤来做，即先进行实体识别，然后抽取实体之间的关系，然而这样做效率低，实体识别的误差会传递到关系抽取部分，导致最终的结果精度较低。

[0050] 为了提高信息抽取的效率和精度，本申请一实施例提供了一种信息抽取方法，参照图1，该方法可以包括：

[0051] 步骤101：获取源数据。

[0052] 在具体实现中，源数据可以为非结构化文本等。

[0053] 步骤102：将源数据输入预先训练得到的信息抽取模型，得到源数据所包含的三元组信息，三元组信息包括源数据中的主语、谓语和宾语；其中，信息抽取模型包括二分类子模型和多标签分类子模型，二分类子模型用于抽取源数据中的主语，多标签分类子模型用于根据主语和源数据，抽取源数据中与主语对应的谓语和宾语。

[0054] 在实际应用中，领域知识图谱通常采用自上而下的方法进行构建，即首先进行顶层设计：确定知识图谱需包含的实体、属性和关系种类。这一部分没有固定标准，通常根据业务需求来进行设计。例如，在艺术领域，我们可能需要获取画作、画家、艺术机构等实体，这些实体和实体之间存在很多属性值和关系，画作存在创作时间、创作媒介等属性，而画家与画作中间存在创作关系等，基于此，我们可以构建如下信息抽取schema：

[0055] {‘subject’:画作，‘predicate’:创作年份，‘object’:年份，‘subject_type’:art_work，‘object_type’:time}；

[0056] {‘subject’:画作，‘predicate’:创作媒介，‘object’:媒介，‘subject_type’:art_work，‘object_type’:medium}；

[0057] {‘subject’:画作，‘predicate’:收藏地，‘object’:艺术机构，‘subject_type’:art_work，‘object_type’:institution}；……

[0058] 其中，subject代表三元组中的主语s，predicate代表三元组中的谓语p也称关系，object代表三元组中的宾语o，subject_type为主语的实体类型，object_type为宾语的实体类型。每一条关系确定一条schema信息，主谓宾确定后，其中的主语s和宾语o的实体也就确定，所以上述schema可以简化为：(画作，创作年份，年份)，(画作，创作媒介，媒介)，(画作，收藏地，技术机构)，……。

[0059] 当源数据为“《蒙娜丽莎》是意大利文艺复兴时期画家达芬奇创作的油画，现收藏于法国卢浮宫博馆”时，该非结构化文本中存在多个三元组：(蒙娜丽莎，作者，达芬奇)、(蒙娜丽莎，收藏地，法国卢浮宫博物馆)、(达芬奇，国籍，意大利)以及(蒙娜丽莎，创作类别，油画)，并且存在一个主语对应多个不同宾语的情况，采用传统的管道式抽取无法同时提取句子中多个三元组。本实施例中我们采用条件概率的思想，首先预测主语s，然后传入主语s来预测该主语s对应的宾语o，然后再传入主语s和宾语o来预测关系谓语p，实际应用中，可以把宾语o和宾语p的预测合并为一步，即：先预测主语s，然后传入主语s来预测该主语s所对应的宾语o及谓语p，如以下公式所示：

[0060] $P(s,p,o)=P(s)P(o|s)P(p|s,o)$

[0061] 其中，信息抽取模型中的二分类子模型和多标签分类子模型可以采用标注好三元组信息的非结构化文本对预训练语言模型和神经网络模型进行联合训练得到。后续实施例中会对信息抽取模型的训练过程以及对非结构化文本进行标注的过程进行详细介绍。

[0062] 在具体实现中，首先将源数据输入二分类子模型，由二分类子模型抽取源数据中的所有主语，然后再将各主语和源数据成对送入多标签分类子模型，由多标签分类子模型抽取源数据中与主语对应的谓语和宾语。这样，只需要将源数据输入信息抽取模型，经信息抽取模型中的二分类子模型和多标签分类子模型处理，就可以输出信息源数据中的三元组信息，即通过端到端的实体和关系联合抽取模型，替代传统的实体识别和关系抽取管道式抽取方法，提高信息抽取的效率和准确率。

[0063] 为了获取信息抽取模型，在一种可选的实现方式中，在步骤102之前还可以包括：获得信息抽取模型的步骤。参照图2，获得信息抽取模型的步骤具体可以包括：

[0064] 步骤201：获得样本集合，样本集合中包括多个待训练文本以及各待训练文本的三元组标注信息，三元组标注信息包括主语标注信息、谓语标注信息和宾语标注信息。

[0065] 其中，待训练文本例如可以为：“《蒙娜丽莎》是意大利文艺复兴时期画家达芬奇创作的油画，现收藏于法国卢浮宫博馆”，该待训练文本的三元组信息包括(蒙娜丽莎，作者，达芬奇)、(蒙娜丽莎，收藏地，法国卢浮宫博物馆)、(达芬奇，国籍，意大利)以及(蒙娜丽莎，创作类别，油画)。

[0066] 在具体实现中，在将待训练文本及对应的三元组信息喂入模型之前，可以按照特定的格式标注三元组信息。具体地，在预测主语时，我们标出主语S在句子中的起止位置。例如，在标注(蒙娜丽莎，作者，达芬奇)，(蒙娜丽莎，创作类别，油画)，(达芬奇，国籍，意大利)时，会将主语蒙娜丽莎和达芬奇在句子中的起止位置分别用两个序列标注出来，即在相应的起始和终止位置标1，其他位置标0，参照图3示出了上述待训练文本的主语标注信息。在预测时，我们可以通过二分类(区分0和1)便可以确定主语的起止位置。在得到主语后，我们利用得到的主语去预测关系(谓语)和宾语，宾语的标注方式和主语相似，区别是，我们会在宾语的起止位置标上谓语对应的索引ID，我们可以预先为每一个谓语建立一个索引，如{1：收藏地，2：作者，3：创作类别，4：国籍，……}，参照图3示出了上述待训练文本的谓语和宾语标注信息。在预测谓语和宾语时我们只需要做一个多标签分类即可。

[0067] 步骤202：将待训练文本输入第一预训练语言模型，将第一预训练语言模型的输出信息送入第一神经网络模型。

[0068] 步骤203：将第一神经网络模型的输出信息以及待训练文本输入第二预训练语言模型，将第二预训练语言模型的输出信息送入第二神经网络模型。

[0069] 步骤204：根据第一神经网络模型的输出信息、第二神经网络模型的输出信息以及三元组标注信息，对第一预训练语言模型、第一神经网络模型、第二预训练语言模型以及第二神经网络模型进行训练，得到信息抽取模型，其中，训练后的第一预训练语言模型和第一神经网络模型构成二分类子模型，训练后的第二预训练语言模型和第二神经网络模型构成多标签分类子模型。

[0070] 在具体实现中，可以根据第一神经网络模型的输出信息以及主语标注信息，确定第一损失函数；根据第二神经网络模型的输出信息、谓语标注信息以及宾语标注信息，确定第二损失函数；对第一预训练语言模型、第一神经网络模型、第二预训练语言模型以及第二

神经网络模型中的参数进行优化，得到信息抽取模型，使得第一损失函数与第二损失函数之和最小。

[0071] 其中，第一预训练语言模型和第二预训练语言模型可以为BERT模型、ERNIE模型以及Span BERT模型等等。下面以第一预训练语言模型和第二预训练语言模型均为BERT模型为例进行说明，第一神经网络模型为Dense层+sigmod，第二神经网络模型为Dense层+softmax，第一损失函数和第二损失函数均为交叉熵损失函数。需要说明的是，第一损失函数和第二损失函数之和最小并不仅限于一个数值，而是一个数值范围。

[0072] 参照图4示出了信息抽取模型的训练框架。模型训练的具体步骤为：首先将待训练文本X，即[CLS]《蒙娜丽莎》是意大利文艺复兴时期画家达芬奇创作的油画……[SEP]，用单输入方式送入BERT模型，将BERT模型输出信息的编码送入Dense层+sigmod，用第一损失函数loss_s(交叉熵损失函数)做二分类训练预测主语起止位置的标注模型，训练后的第一预训练语言模型(BERT)和第一神经网络模型(Dense层+sigmod)构成二分类子模型subject_model。然后随机选取一个主语，如蒙娜丽莎，将其和待训练文本组合成句子对Y，采用双输入方式成对送入BERT模型，如[CLS]《蒙娜丽莎》是意大利文艺复兴时期画家达芬奇创作的油画[SEP]蒙娜丽莎[SEP]，其中[CLS]为分类用的特殊标记位，它表示文本进过BERT后的向量表示，[SEP]为句子间分隔符。我们将BERT模型的输出信息即[CLS]对应的向量送入Dense层+softmax，用第二损失函数loss_o(交叉熵损失函数)做预测谓语和宾语的多分类训练，训练后的第二预训练语言模型(BERT)和第二神经网络模型(Dense层+softmax)构成多标签分类子模型object_model。在实际应用中，可以对二分类子模型subject_model和多标签分类子模型object_model进行联合训练，联合训练的目标是最小化联合损失函数loss＝loss_s+loss_o，对第一预训练语言模型、第一神经网络模型、第二预训练语言模型以及第二神经网络模型中的参数进行迭代优化，从而得到信息抽取模型。

[0073] 具体的，在主语抽取任务上，输入样本X经BERT编码后的输出信息可以表示为：

[0074] $h_0＝XW_t+W_p$

[0075] $h_l＝Transformer(l-1)\ l∈[1,L]$ (1)

[0076] 其中W_t为词嵌入矩阵，h_i隐藏层向量(即第i层Transformer网络的输出)，L表示Transformer的层数。

[0077] 这里我们采用两个二分类来判断输入序列在当前位置上是0/1的可能性来确定主语的起止位置，即通过主语起始位置序列S_s和主语终止位置序列S_e中每个位置上可能为一个主语起止位置的置信度来确定一个主语，如某个主语的起始位置可能在S_s中每个位置上出现的概率分布p_i^{s-s}(置信度)可以表示为：

[0078] $$p_i^{s-s}=\sigma(S_s),\ S_s=W_{start}h_l^i+b_{start} \quad (2)$$

[0079] 其中W_{start}为可训练权重向量，b_{start}为偏置项，σ为sigmoid激活函数，h_l^i为第i个输入序列经过BERT后的编码表示，由(1)式获得。同理，其终止位置在S_e中每个位置出现的概率分布p_i^{s-e}可以表示为：

[0080] $$p_i^{s-e}=\sigma(S_e),\ S_e=W_{end}h_l^i+b_{end} \quad (3)$$

[0081] 最终我们得到两个向量p_i^{s-s}，p_i^{s-e}，训练的目标函数为：

[0082] $$loss\_s = -\sum_{i=1}^{n+2} p_i^{s-s} \log(p_i^{s-s}) - \sum_{j=1}^{n+2} p_j^{s-e} \log(p_j^{s-e}) \quad (4)$$

[0083] 同理，在进行宾语和关系(谓语)抽取时，我们从主语中随机采样一个主语，将其与句子组合成句子对嵌入的方式，用BERT进行编码得到编码表示：

[0084] $h_0 = YW_t + W_s + W_p$

[0085] $h_l = Transformer(h_{l-1}), l \in [1, L]$ (5)

[0086] 其中W_s为句子嵌入矩阵。

[0087] 我们同样用两个序列来确定宾语的起止位置，如图3所示，与主语抽取方式不同的是，我们用多标签分类的方式同时确定宾语的起止的位置和关系，即在宾语的起止位置上确定关系标签的概率p_i^{o-r-s}， p_i^{o-r-e} ：

[0088] $$p_i^{o-r-s} = \alpha(\tilde{S}_s),\ \tilde{S}_s = W_{start}^r h_l^i + b_{start}^r \quad (6)$$

[0089] $$p_i^{o-r-e} = \alpha(\tilde{S}_e),\ \tilde{S}_e = W_{end}^r h_l^i + b_{end}^r \quad (7)$$

[0090] 其中W_{start}^r为可训练权重向量，b_{start}^r为偏置项，α为softmax激活函数。训练的目标函数为：

[0091] $$loss\_o = -\sum_{i=1}^{n+2}\sum_{r}^{R} t_{r,i}^s \log(p_i^{o-r-s}) - \sum_{j=1}^{n+2}\sum_{r}^{R} t_{r,j}^e \log(p_i^{o-r-e})$$

[0092] 其中，$t_{r,i}^s$为真实的关系标签，R为关系标签的数量。

[0093] 模型训练过程中待优化的参数为上述的可训练权重向量，通过对参数进行迭代更新优化，使损失函数loss最小化。

[0094] 目前主流的关系抽取方法是有监督的学习方法、半监督的学习方法和无监督的学习方法三种。与半监督的学习方法和无监督的学习方法相比，有监督的学习方法准确率与召回率更高，因此受到越来越多的关注。有监督的学习方法需要大量的数据标注，如果提高数据标注效率也是一个急需解决的问题。

[0095] 为了提高数据标注效率，在一种可选的实现方式中，参照图5，步骤201可以包括：

[0096] 步骤501：对非结构化文本样本进行处理，得到待标注文本。

[0097] 步骤502：获取已完成标注的待训练文本以及待训练文本的三元组标注信息。

[0098] 步骤503：响应于待标注文本中包含三元组标注信息中的主语标注信息和宾语标注信息，按照三元组标注信息对待标注文本进行标注。

[0099] 采用有监督方法进行信息抽取需要大量的标注数据，这需要消耗大量的人力和财力成本进行数据标注。当具有一定规模的知识库时，我们可以采用远程监督的方法进行语料的自动化标注，在此基础上进行人工审核，处理错标和漏标问题。参照图6示出了自动化标注的流程图，其中的非结构化数据可以从艺术领域网站爬取，也可以从当前知识图谱中的非结构化信息中获取。当然，也可以利用实体词直接从百度百科等搜索引擎中搜索得到。在实际应用中，可以首先对从网页上爬取到的非结构数据进行处理如数据清洗等，从而去除无用的标点符号和脚本等无用信息，得到待标注文本Sentence，然后再利用预先定义好的schema和知识图谱中的三元组标注信息进行远程监督方式标注。

[0100] 在具体实现中，可以判断待标注文本Sentence中是否存在现有知识图谱中的三元组标注信息中的主语e1和宾语e2，如果二者同时存在，则按照现有知识图谱中的三元组标注信息对待标注文本进行标注。这样，通过利用已有知识库自动化标注数据，可以减轻语料标注的成本。

[0101] 当待标注文本为“《蒙娜丽莎》是意大利文艺复兴时期画家达芬奇创作的油画，现收藏于法国卢浮宫博物馆”时，该待标注文本的标注格式如下：

[0102] {’text’:‘《蒙娜丽莎》是意大利文艺复兴时期画家达芬奇创作的油画，现收藏于法国卢浮宫博物馆’,‘spo_list’:[(蒙娜丽莎，作者，达芬奇)，(蒙娜丽莎，收藏地，法国卢浮宫博物馆)，(达芬奇，国籍，意大利)，(蒙娜丽莎，创作类别，油画)]}。

[0103] 另外，为了尽可能降低训练数据中的噪音和漏标数据，我们可以采用知识蒸馏的方法对自动标注数据进行降噪。上述实现方式还可以包括：

[0104] 步骤504：采用预先训练得到的K个预测模型对待标注文本进行预测，得到K个三元组预测信息。

[0105] 其中，K个预测模型可以根据已完成标注的待训练文本以及待训练文本的三元组标注信息，采用K折交叉验证的方式训练得到。

[0106] 具体为：将训练样本等分为K份，依次取其中的K-1份训练模型，另外1份作为待预测样本。如可分为[D1,D2,D3,…,DK]，依次取[D1,D2,…,Dk-1,Dk+1,…,DK]为训练样本，Dk为待预测样本，k∈[1,K]。

[0107] 步骤505：当第一三元组信息的数量与K的比值大于第一预设阈值时，将第一三元组信息作为待标注文本的三元组标注信息添加至样本集合中，其中，第一三元组信息为出现在三元组预测信息中但未出现在待标注文本的三元组标注信息中的三元组信息。

[0108] 步骤506：当第二三元组信息的数量与K的比值大于第二预设阈值时，将第二三元组信息从待标注文本的三元组标注信息中删除，其中，第二三元组信息为出现在待标注文本的三元组标注信息中但未出现在三元组预测信息中的三元组信息。

[0109] 其中，K值可以大于或等于5且小于或等于10，也可以依据数据规模自行设定。第一预设阈值和第二预设阈值可以相同或不同，具体数值可以根据实际需求确定。

[0110] 在具体实现中，可以采用K折交叉验证的方式用已标注数据训练出K个模型，然后用训练好的K个模型去预测待标注文本。模型预测出来的结果和原始标注结果会有偏差，例如在某个待标注文本S中，被标注出了{T1,T2,T3,…Tk}K个三元组标注信息，记R_s＝{T1,T2,T3,…Tk}，然后可以用K个模型去预测待标注文本S，得到K个三元组预测信息。K个三元组预测信息中可能存在某个第一三元组信息Ti不在R_s中，该第一三元组信息Ti在K个三元组预测信息中出现了M次，K个三元组预测信息中可能有N个结果不包含第二三元组信息Tj，而第二三元组信息Tj存在于R_s中。此时，我们可以设置第一预设阈值和第二预设阈值均为Score，当M/K>Score时，认为第一三元组信息Ti为待标注文本的漏标数据，因此可以将第一三元组信息Ti添加到三元组标注信息R_s中，当N/K>Score时，认为第二三元组信息Tj为错标数据，因此，需要将第二三元组信息Tj从三元组标注信息R_s中删除。按照此方式重复训练和预测多次，可以不断修正训练样本集合。

[0111] 本实现方式中，利用已有知识库自动化标注数据，从而可以降低预料标注的成本，在此基础上进行人工审核，并在后期利用知识蒸馏的方法对标注的数据进行降噪处理。

[0112] 本实施例提供的信息抽取方法，参照图7，主要涉及数据标注方法、schema构建、信息抽取算法模型、数据降噪等几个主要部分，该方案运用端到端的实体关系联合抽取方法从非结构化文本中抽取知识，在保证信息抽取精度的同时，降低构建知识图谱的代价，提升信息抽取效率，节约人力成本。

[0113] 本申请另一实施例还提供了一种信息抽取装置，参照图8，该装置可以包括：

[0114] 获取模块801，被配置为获取源数据；

[0115] 抽取模块802，被配置为将所述源数据输入预先训练得到的信息抽取模型，得到所述源数据所包含的三元组信息，所述三元组信息包括所述源数据中的主语、谓语和宾语；其中，所述信息抽取模型包括二分类子模型和多标签分类子模型，所述二分类子模型用于抽取所述源数据中的主语，所述多标签分类子模型用于根据所述主语和所述源数据，抽取所述源数据中与所述主语对应的谓语和宾语。

[0116] 在一种可选的实现方式中，所述装置还可以包括：模型获取模块，被配置为获得所述信息抽取模型，所述模型获取模块包括：

[0117] 第一单元，被配置为获得样本集合，所述样本集合中包括多个待训练文本以及各所述待训练文本的三元组标注信息，所述三元组标注信息包括主语标注信息、谓语标注信息和宾语标注信息；

[0118] 第二单元，被配置为将所述待训练文本输入第一预训练语言模型，将所述第一预训练语言模型的输出信息送入第一神经网络模型；

[0119] 第三单元，被配置为将所述第一神经网络模型的输出信息以及所述待训练文本输入第二预训练语言模型，将所述第二预训练语言模型的输出信息送入第二神经网络模型；

[0120] 第四单元，被配置为根据所述第一神经网络模型的输出信息、所述第二神经网络模型的输出信息以及所述三元组标注信息，对所述第一预训练语言模型、所述第一神经网络模型、所述第二预训练语言模型以及所述第二神经网络模型进行训练，得到所述信息抽取模型，其中，训练后的第一预训练语言模型和第一神经网络模型构成所述二分类子模型，训练后的第二预训练语言模型和第二神经网络模型构成所述多标签分类子模型。

[0121] 在一种可选的实现方式中，所述第四单元具体被配置为：

[0122] 根据所述第一神经网络模型的输出信息以及所述主语标注信息，确定第一损失函数；

[0123] 根据所述第二神经网络模型的输出信息、所述谓语标注信息以及所述宾语标注信息，确定第二损失函数；

[0124] 对所述第一预训练语言模型、所述第一神经网络模型、所述第二预训练语言模型以及所述第二神经网络模型中的参数进行优化，得到所述信息抽取模型，使得所述第一损失函数与所述第二损失函数之和最小。

[0125] 在一种可选的实现方式中，所述第一损失函数和所述第二损失函数均为交叉熵损失函数。

[0126] 在一种可选的实现方式中，所述第一单元具体被配置为：

[0127] 获取非结构化文本样本；

[0128] 对所述非结构化文本样本进行处理，得到待标注文本；

[0129] 获取已完成标注的待训练文本以及所述待训练文本的三元组标注信息；

[0130] 响应于所述待标注文本中包含所述三元组标注信息中的主语标注信息和宾语标注信息，按照所述三元组标注信息对所述待标注文本进行标注。

[0131] 在一种可选的实现方式中，所述第一单元还被配置为：

[0132] 采用预先训练得到的K个预测模型对所述待标注文本进行预测，得到K个三元组预测信息；

[0133] 当第一三元组信息的数量与K的比值大于第一预设阈值时，将所述第一三元组信息作为所述待标注文本的三元组标注信息添加至所述样本集合中，其中，所述第一三元组信息为出现在所述三元组预测信息中但未出现在所述待标注文本的三元组标注信息中的三元组信息；

[0134] 当第二三元组信息的数量与K的比值大于第二预设阈值时，将所述第二三元组信息从所述待标注文本的三元组标注信息中删除，其中，所述第二三元组信息为出现在所述待标注文本的三元组标注信息中但未出现在所述三元组预测信息中的三元组信息。

[0135] 其中，K值可以大于或等于5且小于或等于10，也可以依据数据规模自行设定。

[0136] 在一种可选的实现方式中，所述第一单元还被配置为：

[0137] 根据已完成标注的待训练文本以及所述待训练文本的三元组标注信息，采用K折交叉验证的方式获得K个预测模型。

[0138] 关于上述实施例中的装置，其中各个模块执行操作的具体方式已经在应用于服务器的信息抽取方法的实施例中进行了详细描述，此处将不做详细阐述说明。

[0139] 本申请另一实施例还提供了一种电子设备，该电子设备包括：

[0140] 处理器；

[0141] 用于存储所述处理器可执行指令的存储器；

[0142] 其中，所述处理器被配置为执行所述指令，以实现任一实施例所述的信息抽取方法。

[0143] 本申请另一实施例还提供了一种存储介质，当所述存储介质中的指令由电子设备的处理器执行时，使得所述电子设备能够执行任一实施例所述的信息抽取方法。

[0144] 本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。

[0145] 最后，还需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅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”或者其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、商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、商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、商品或者设备中还存在另外的相同要素。

[0146] 以上对本发明所提供的一种信息抽取方法、装置、电子设备及存储介质进行了详细介绍，本文中应用了具体个例对本发明的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本发明的方法及其核心思想；同时，对于本领域的一般技术人员，依据本发明的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本发明的限制。

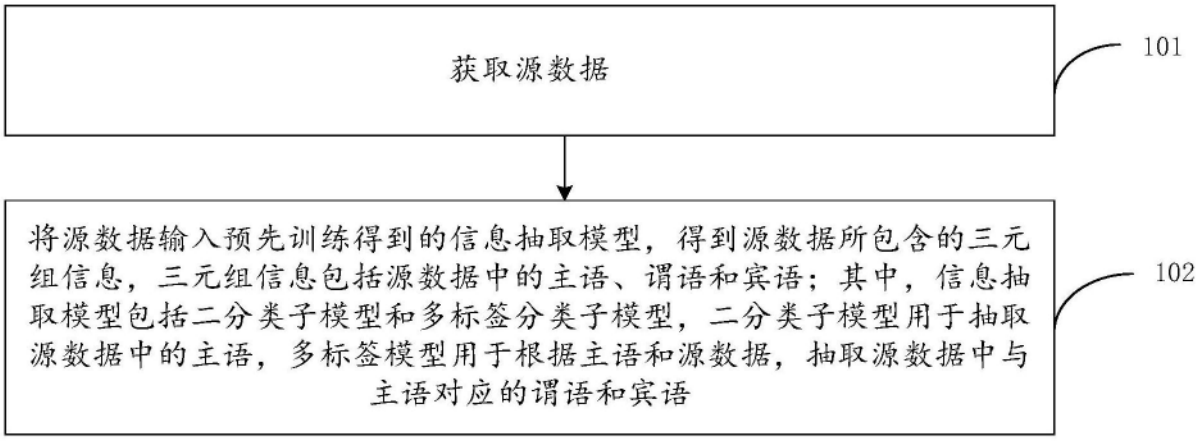

获取源数据
101
将源数据输入预先训练得到的信息抽取模型，得到源数据所包含的三元组信息，三元组信息包括源数据中的主语、谓语和宾语；其中，信息抽取模型包括二分类子模型和多标签分类子模型，二分类子模型用于抽取源数据中的主语，多标签模型用于根据主语和源数据，抽取源数据中与主语对应的谓语和宾语
102

图1

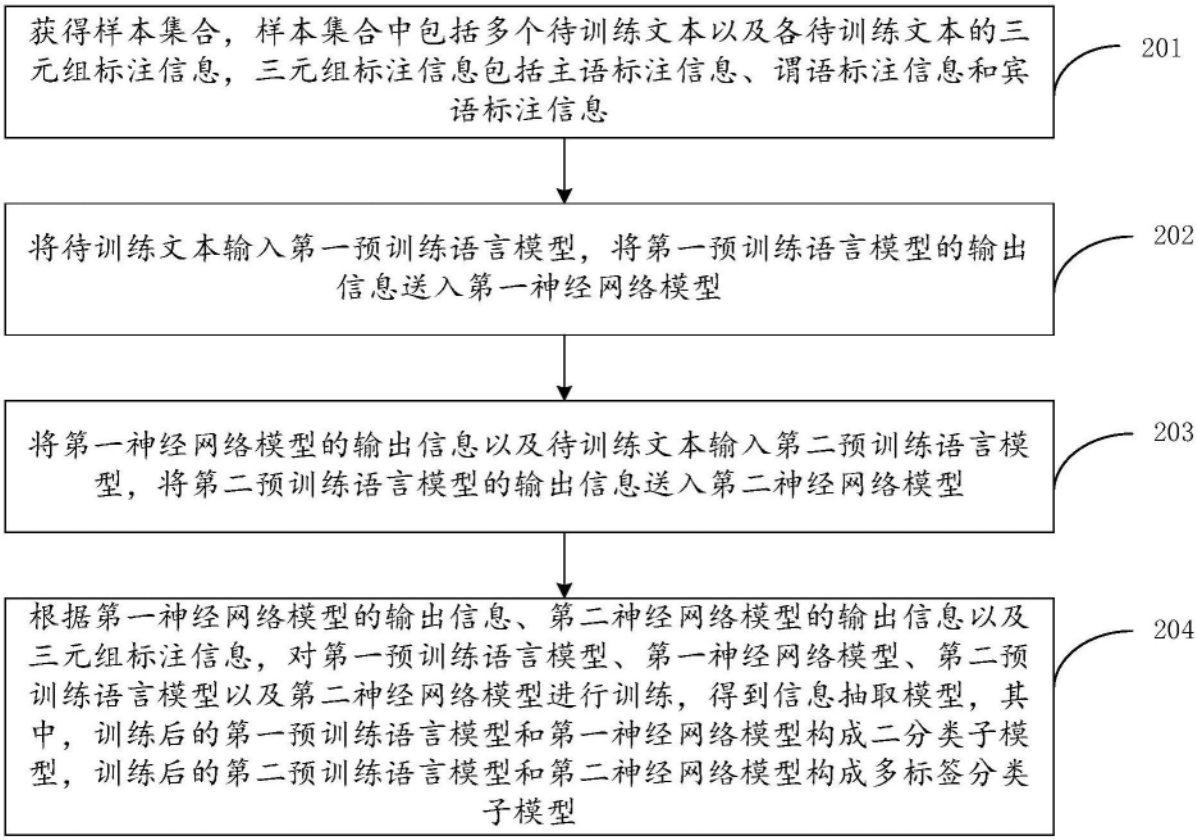

获得样本集合，样本集合中包括多个待训练文本以及各待训练文本的三元组标注信息，三元组标注信息包括主语标注信息、谓语标注信息和宾语标注信息
201
将待训练文本输入第一预训练语言模型，将第一预训练语言模型的输出信息送入第一神经网络模型
202
将第一神经网络模型的输出信息以及待训练文本输入第二预训练语言模型，将第二预训练语言模型的输出信息送入第二神经网络模型
203
根据第一神经网络模型的输出信息、第二神经网络模型的输出信息以及三元组标注信息，对第一预训练语言模型、第一神经网络模型、第二预训练语言模型以及第二神经网络模型进行训练，得到信息抽取模型，其中，训练后的第一预训练语言模型和第一神经网络模型构成二分类子模型，训练后的第二预训练语言模型和第二神经网络模型构成多标签分类子模型
204

图2

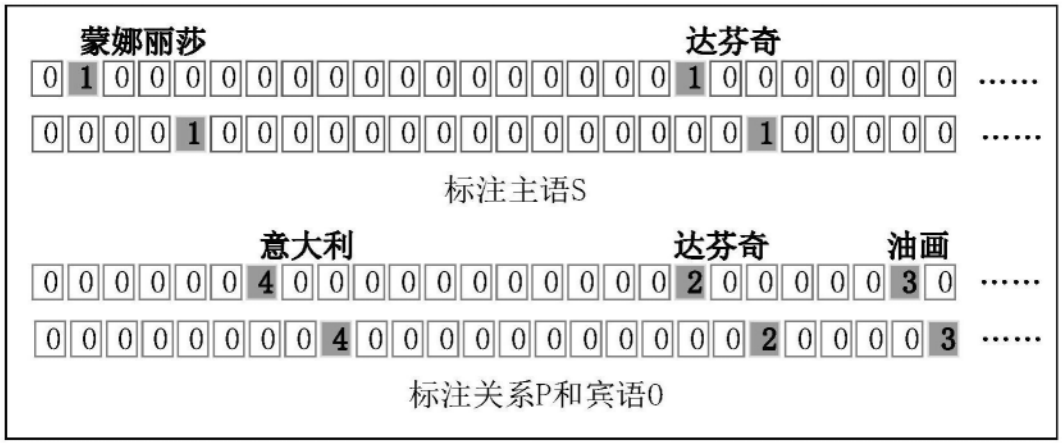

蒙娜丽莎
达芬奇
标注主语S
意大利
达芬奇
油画
标注关系P和宾语O

图3

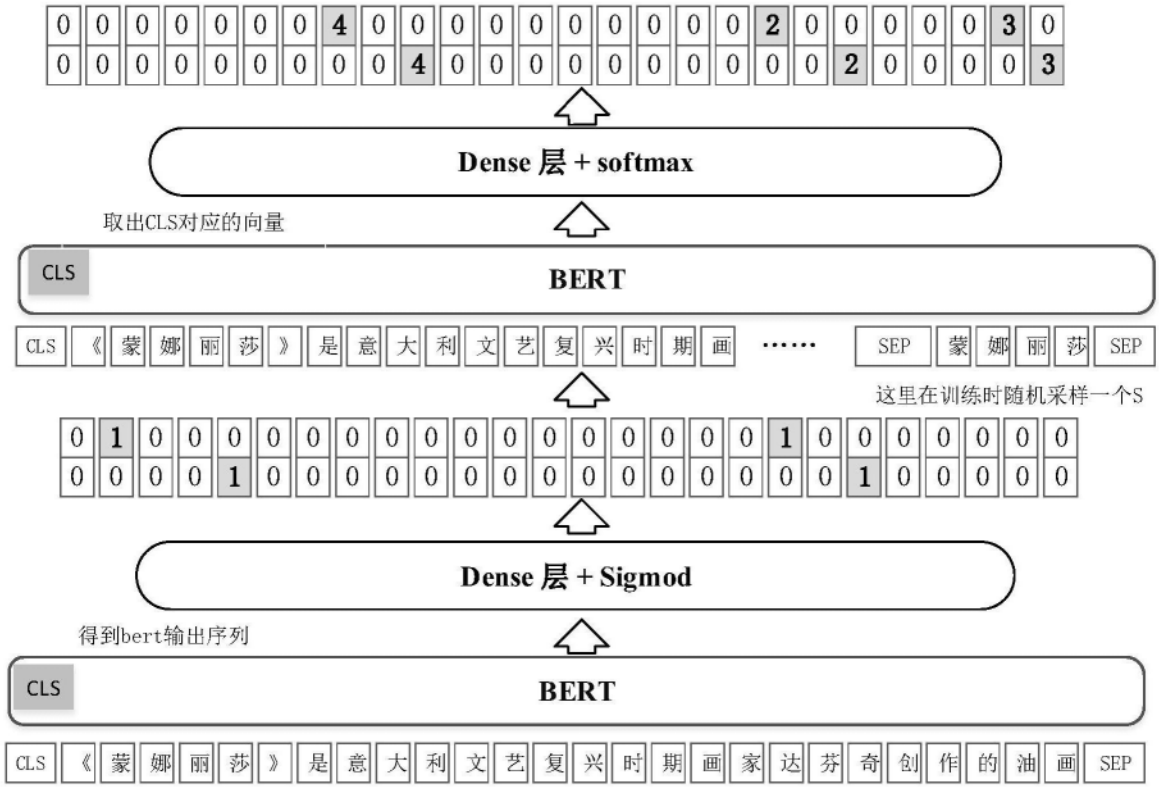

Dense 层 + softmax
取出CLS对应的向量
CLS
BERT
CLS 《 蒙 娜 丽 莎 》 是 意 大 利 文 艺 复 兴 时 期 画 …… SEP 蒙 娜 丽 莎 SEP
这里在训练时随机采样一个S
Dense 层 + Sigmod
得到bert输出序列
CLS
BERT
CLS 《 蒙 娜 丽 莎 》 是 意 大 利 文 艺 复 兴 时 期 画 家 达 芬 奇 创 作 的 油 画 SEP

图4

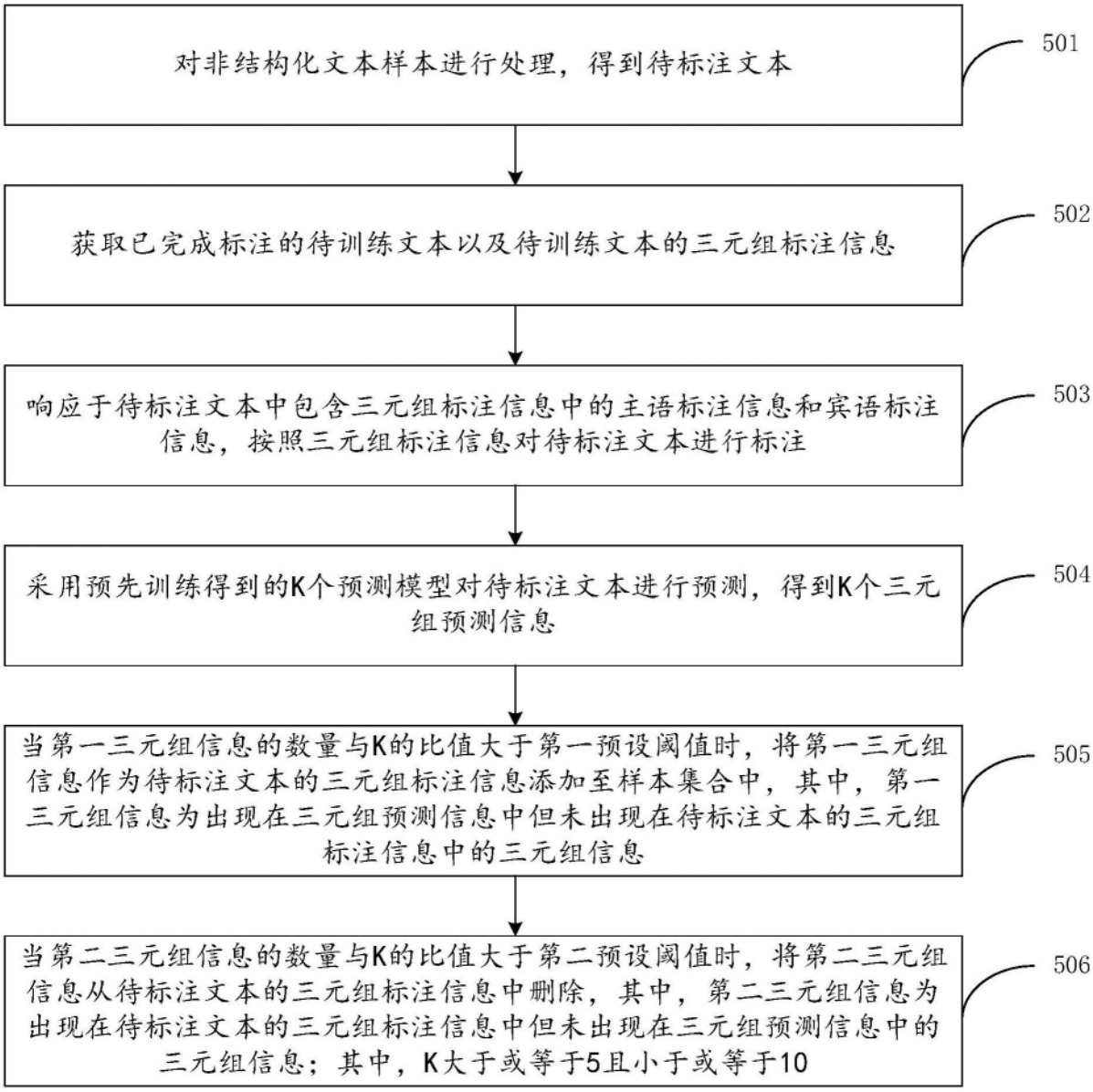

对非结构化文本样本进行处理，得到待标注文本
501
获取已完成标注的待训练文本以及待训练文本的三元组标注信息
502
响应于待标注文本中包含三元组标注信息中的主语标注信息和宾语标注信息，按照三元组标注信息对待标注文本进行标注
503
采用预先训练得到的K个预测模型对待标注文本进行预测，得到K个三元组预测信息
504
当第一三元组信息的数量与K的比值大于第一预设阈值时，将第一三元组信息作为待标注文本的三元组标注信息添加至样本集合中，其中，第一三元组信息为出现在三元组预测信息中但未出现在待标注文本的三元组标注信息中的三元组信息
505
当第二三元组信息的数量与K的比值大于第二预设阈值时，将第二三元组信息从待标注文本的三元组标注信息中删除，其中，第二三元组信息为出现在待标注文本的三元组标注信息中但未出现在三元组预测信息中的三元组信息；其中，K大于或等于5且小于或等于10
506

图5

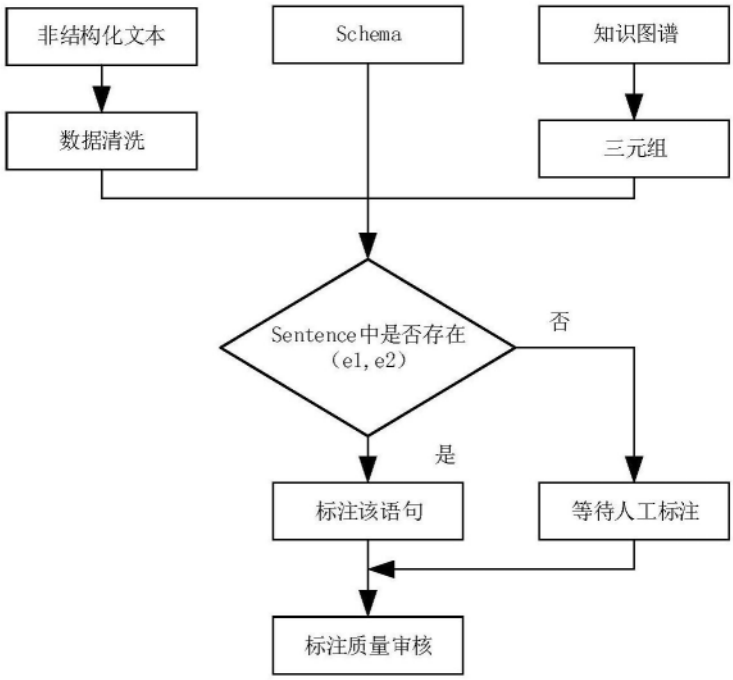

非结构化文本
Schema
知识图谱
数据清洗
三元组
Sentence中是否存在
(e1,e2)
否
是
标注该语句
等待人工标注
标注质量审核

图6

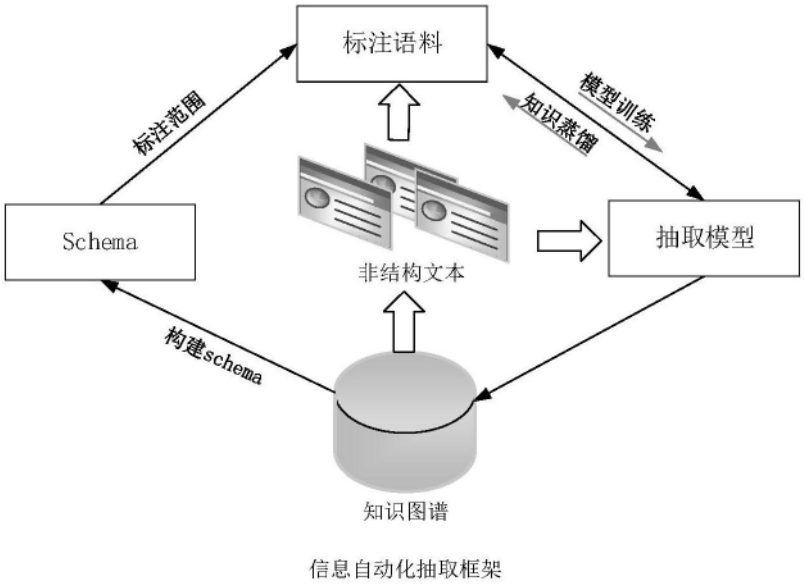

标注语料
标注范围
模型训练
知识蒸馏
Schema
抽取模型
非结构文本
构建schema
知识图谱
信息自动化抽取框架

图7

图8