(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112364174 A
(43)申请公布日 2021.02.12

(21)申请号 202011131273.1

(22)申请日 2020.10.21

(71)申请人 山东大学
地址 250100 山东省济南市历城区山大南路27号

(72)发明人 郭伟 宋贤 鹿旭东 孔兰菊 崔立真

(74)专利代理机构 济南圣达知识产权代理有限公司 37221
代理人 祖之强

(51)Int.Cl.
G06F 16/36(2019.01)
G06F 40/295(2020.01)
G06F 40/30(2020.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)
G16H 10/60(2018.01)

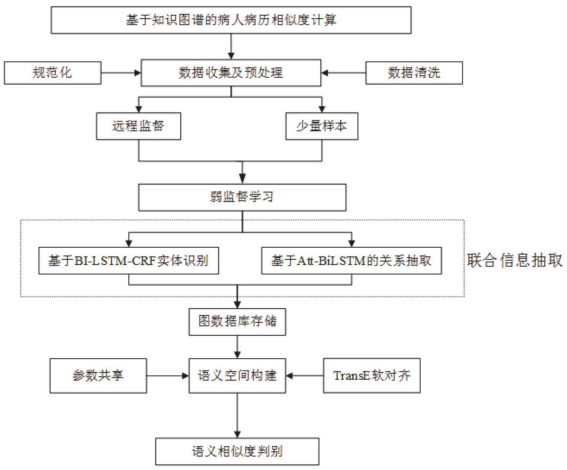
权利要求书2页 说明书8页 附图3页

(54)发明名称

基于知识图谱的病人病历相似度评估方法及系统

(57)摘要

本发明提供了一种基于知识图谱的病人病历相似度评估方法及系统，获取病人病历文本数据并进行预处理；采用基于弱监督学习的联合信息抽取模型对预处理后的数据进行实体识别和实体关系抽取，将得到的医学实体类型用知识向量进行表示；根据得到的实体关系进行知识图谱中三元组的构建，并使用联合知识向量进行表示；利用联合向量将知识向量合并到同一个语义空间，通过计算实体之间的语义距离进行相似度的判别；本发明通过多次使用双向循环神经网络来挖掘抽取电子病历中的重要知识，将学科领域知识图谱概念扩展到医疗领域中，在医学实体类型相对有限的现状下进行实体识别和关系抽取，提高了相似度评估的准确度。



CN 112364174 A

1.一种基于知识图谱的病人病历相似度评估方法,其特征在于,包括以下步骤:

获取病人病历文本数据并进行预处理;

采用基于弱监督学习的联合信息抽取模型对预处理后的数据进行实体识别和实体关系抽取,将得到的医学实体类型用知识向量进行表示;

根据得到的实体关系进行知识图谱中三元组的构建,并使用联合知识向量进行表示;

利用联合向量将知识向量合并到同一个语义空间,通过计算实体之间的语义距离进行相似度的判别。

2.如权利要求1所述的基于知识图谱的病人病历相似度评估方法,其特征在于,联合信息抽取模型中,采用带有条件随机场层的双向LSTM网络进行实体识别,将得到的医学实体类型用知识向量进行表示。

3.如权利要求1所述的基于知识图谱的病人病历相似度评估方法,其特征在于,联合信息抽取模型中,基于注意力机制的带有条件随机场层的双向LSTM网络进行实体关系抽取。

4.如权利要求3所述的基于知识图谱的病人病历相似度评估方法,其特征在于,基于注意力机制的带有条件随机场层的双向LSTM网络,包括输入层、Embedding层、LSTM层、Attention层和输出层;

Embedding层,被配置为:将句子中的每一个词映射成固定长度的向量;

LSTM层,被配置为:利用双向的LSTM对embedding向量计算;

Attention层,被配置为:对双向LSTM的结果使用Attention加权。

5.如权利要求1所述的基于知识图谱的病人病历相似度评估方法,其特征在于,相似度的判断,具体为:

利用联合向量将知识向量合并到同一个语义空间,将多个知识图谱的三元组糅合在一起共同训练,对视为具有相似关系的三元组所在的空间进行约束;

通过带参数共享和软对齐的TransE模型进行联合知识嵌入,计算实体之间的语义距离实现相似度的判别。

6.如权利要求5所述的基于知识图谱的病人病历相似度评估方法,其特征在于,通过知识表示学习的方式将原始数据中涉及到的实体和关系进行再训练,转换为嵌入向量形式;

采用TransE模型进行再训练,将训练集随机初始化为向量的形式作为输入,并产生训练集中实体集和预定义的关系集所对应的词向量作为输出;

给定实体集、关系集和训练集,通过训练集随机的替换头实体或者尾实体以构建负样本,计算正确三元组实体和关系之间的距离,负样本中实体关系的距离,并调整两者之间的误差,将实体关系表示成符合现实关系的向量。

7.如权利要求5所述的基于知识图谱的病人病历相似度评估方法,其特征在于,在合并的语义空间通过实体之间的语义距离实现实体之间的对齐,使用新对齐得到的实体对更新联合向量和找到新的实体对。

8.一种基于知识图谱的病人病历相似度评估系统,其特征在于,包括:

数据获取模块,被配置为:获取病人病历文本数据并进行预处理;

信息抽取模块,被配置为:采用基于弱监督学习的联合信息抽取模型对预处理后的数据进行实体识别和实体关系抽取,将得到的医学实体类型用知识向量进行表示;

三元组构建模块,被配置为:根据得到的实体关系进行知识图谱中三元组的构建,并使

用联合知识向量进行表示；

相似度判别模块，被配置为：利用联合向量将知识向量合并到同一个语义空间，通过计算实体之间的语义距离进行相似度的判别。

9.一种计算机可读存储介质，其上存储有程序，其特征在于，该程序被处理器执行时实现如权利要求1-7任一项所述的基于知识图谱的病人病历相似度评估方法中的步骤。

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1-7任一项所述的基于知识图谱的病人病历相似度评估方法中的步骤。

基于知识图谱的病人病历相似度评估方法及系统

技术领域

[0001] 本发明涉及文本数据处理技术领域，特别涉及一种基于知识图谱的病人病历相似度评估方法及系统。

背景技术

[0002] 本部分的陈述仅仅是提供了与本发明相关的背景技术，并不必然构成现有技术。

[0003] 电子病历(electronic medical record，EMR)是指医务人员在对患者医疗的过程中，使用医疗机构信息系统生成的文字符号、图表、图形、数据等数字化电子信息，具有存储、管理、传输和重现医疗记录的作用。随着智慧医疗和信息化的不断发展，电子病例逐渐取代纸质病历，成为记录个人医疗信息、健康信息的主要载体，它对患者的医疗活动有比较详细的记录，包含了大量与患者健康状况相关的医疗信息，对于医疗领域中的诊断、案例分析、预后等潜藏着巨大的利用价值。面对海量、信息种类多样化的病例数据，通过对非结构化的电子病历进行相似性查重与分析，辅助查找出违规拷贝的病历内容，可以有效地控制医院病历质量。将非结构化的病历文本进行信息抽取，从而使其结构化的方法中又分为两种，一种是基于规则的，即设计符合一定语法规则的正则表达式，用正则表达式到病历里进行匹配，找到相应的信息类型，比如疾病、症状、治疗手段和检查手段等。另一种是基于自然语言处理的，主要依赖于分词、信息抽取、句法分析等技术抽取出非结构化文本中的实体词，然后与医学术语词典中的词进行匹配，得到病历文本中的医学词汇进行实体识别。

[0004] 信息抽取是构建知识图谱的重要步骤，通常被拆分为实体识别和关系抽取两个子任务，在用于解决实体及其关系的提取问题中，Pipelined方法和联合学习法被广泛的应用。Pipelined方法将信息抽取视为两个独立的任务，即命名实体识别(Named Entity Recognition，简称NER)和关系分类(Relation Classification，简称RC)。此外，知识图谱是将信息中的知识要点、内在联系进行可视化形式展示，是实现智慧医疗的基石，对于临床决策支持和个性化医疗服务等具有重要意义。

[0005] 本发明发明人发现，知识图谱在医疗方面还没有得到广泛应用，主要在于非结构化文本抽取和知识图谱绘制这两方面的困难；文本信息抽取分两个子任务实现会产生不可忽略的误差传播，现有的有监督学习存在耗时费力的情况，从而导致信息抽取的三元组信息存在缺失、不准确、不通顺等情况。

发明内容

[0006] 为了解决现有技术的不足，本发明提供了一种基于知识图谱的病人病历相似度评估方法及系统，采用深度学习方法从电子病历中进行信息抽取，通过多次使用双向循环神经网络来挖掘抽取电子病历中的重要知识，将学科领域知识图谱概念扩展到医疗领域中，在医学实体类型相对有限的现状下进行实体识别和关系抽取，提高了相似度评估的准确度。

[0007] 为了实现上述目的，本发明采用如下技术方案：

[0008] 本发明第一方面提供了一种基于知识图谱的病人病历相似度评估方法。

[0009] 一种基于知识图谱的病人病历相似度评估方法，包括以下步骤：

[0010] 获取病人病历文本数据并进行预处理；

[0011] 采用基于弱监督学习的联合信息抽取模型对预处理后的数据进行实体识别和实体关系抽取，将得到的医学实体类型用知识向量进行表示；

[0012] 根据得到的实体关系进行知识图谱中三元组的构建，并使用联合知识向量进行表示；

[0013] 利用联合向量将知识向量合并到同一个语义空间，通过计算疾病实体之间的语义距离进行病历相似度的判别。

[0014] 本发明第二方面提供了一种基于知识图谱的病人病历相似度评估系统。

[0015] 一种基于知识图谱的病人病历相似度评估系统，包括：

[0016] 数据获取模块，被配置为：获取病人病历文本数据并进行预处理；

[0017] 信息抽取模块，被配置为：采用基于弱监督学习的联合信息抽取模型对预处理后的数据进行实体识别和实体关系抽取，将得到的医学实体类型用知识向量进行表示；

[0018] 三元组构建模块，被配置为：根据得到的实体关系进行知识图谱中三元组的构建，并使用联合知识向量进行表示；

[0019] 相似度判别模块，被配置为：利用联合向量将知识向量合并到同一个语义空间，通过计算实体之间的语义距离进行相似度的判别。

[0020] 本发明第三方面提供了一种计算机可读存储介质，其上存储有程序，该程序被处理器执行时实现如本发明第一方面所述的基于知识图谱的病人病历相似度评估方法中的步骤。

[0021] 本发明第四方面提供了一种电子设备，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程序，所述处理器执行所述程序时实现如本发明第一方面所述的基于知识图谱的病人病历相似度评估方法中的步骤。

[0022] 与现有技术相比，本发明的有益效果是：

[0023] 1、本发明所述的方法、系统、介质或电子设备，采用深度学习方法从电子病历中进行信息抽取，通过多次使用双向循环神经网络来挖掘抽取电子病历中的重要知识，将学科领域知识图谱概念扩展到医疗领域中，在医学实体类型相对有限的现状下进行实体识别和关系抽取，提高了相似度评估的准确度。

[0024] 2、本发明所述的方法、系统、介质或电子设备，将非结构化的病历数据进行数据预处理，利用深层次的语义信息进行病历数据的相似度计算，与传统字段全匹配的方法相比，识别准确率大大提高。

[0025] 3、本发明所述的方法、系统、介质或电子设备，在BI-LSTM-CRF网络中加入注意力机制，构建出联合知识抽取网络，在对实体识别的同时也进行关系抽取，用于知识图谱中的三元组的快速构建，避免了将信息抽取分为两个子任务进行导致的子任务之间的误差传播。

[0026] 4、本发明所述的方法、系统、介质或电子设备，基于弱监督学习的联合知识抽取网络对数据进行有效信息抽取的同时，减少了人工标注数据集的繁琐工作，具有较好的可扩展性。

附图说明

[0027] 构成本发明的一部分的说明书附图用来提供对本发明的进一步理解，本发明的示意性实施例及其说明用于解释本发明，并不构成对本发明的不当限定。

[0028] 图1为本发明实施例1提供的基于知识图谱的病人病历相似度评估方法的流程示意图。

[0029] 图2为本发明实施例1提供的BI-LSTM网络结构示意图。

[0030] 图3为本发明实施例1提供的LSTM-CRF网络结构示意图。

[0031] 图4为本发明实施例1提供的BI-LSTM-CRF网络结构示意图。

[0032] 图5为本发明实施例1提供的Att-BiLSTM网络结构示意图。

具体实施方式

[0033] 下面结合附图与实施例对本发明作进一步说明。

[0034] 应该指出，以下详细说明都是示例性的，旨在对本发明提供进一步的说明。除非另有指明，本文使用的所有技术和科学术语具有与本发明所属技术领域的普通技术人员通常理解的相同含义。

[0035] 需要注意的是，这里所使用的术语仅是为了描述具体实施方式，而非意图限制根据本发明的示例性实施方式。如在这里所使用的，除非上下文另外明确指出，否则单数形式也意图包括复数形式，此外，还应当理解的是，当在本说明书中使用术语“包含”和/或“包括”时，其指明存在特征、步骤、操作、器件、组件和/或它们的组合。

[0036] 在不冲突的情况下，本发明中的实施例及实施例中的特征可以相互组合。

[0037] 实施例1：

[0038] 如图1所示，本发明实施例1提供了一种基于知识图谱的病人病历相似度评估方法，包括以下步骤：

[0039] 首先，采用带有条件随机场层的双向LSTM网络(BI-LSTM-CRF)进行实体识别，将得到的医学实体类型用知识向量进行表示；

[0040] 然后在BI-LSTM-CRF网络中加入注意力机制(Att-BiLSTM)，构建出联合知识抽取网络(Joint knowledge extraction network，JKENet)在对实体识别的同时也进行关系抽取，用于知识图谱中的三元组的快速构建，使用联合知识向量进行表示；

[0041] 最后，利用联合向量将知识向量合并到同一个语义空间，通过计算实体之间的语义距离实现相似度的判别。

[0042] 详细的，包括以下内容：

[0043] 本实施例中，在对数据进行有效信息抽取的同时，通过弱监督学习减少了人工标注数据集的繁琐工作，相对监督学习之类的方法，具有较好的可扩展性。

[0044] 通过采用带有条件随机场层的双向LSTM网络(BI-LSTM-CRF)进行医学实体识别，得到医学实体类型并用知识向量进行表示。

[0045] 在BI-LSTM-CRF网络中引入注意力机制，计算注意力概率以突出关键词汇在病历中的重要程度，使用Att-BiLSTM网络模型进行演化关系抽取得到结果。构建出联合知识抽取网络在对实体识别的同时也进行关系抽取，用于知识图谱中的三元组的构建。

[0046] 利用联合向量将知识向量合并到同一个语义空间，将多个知识图谱的三元组糅合

在一起共同训练，对视为具有相似关系的三元组所在的空间进行约束；通过带参数共享和软对齐的TransE实现联合知识嵌入，计算实体之间的语义距离实现相似度的判别。

[0047] (1)基于条件随机场层的双向LSTM网络(BI-LSTM-CRF)

[0048] 双向LSTM网络能够在指定的时间范围内有效地使用过去的特征(通过前向状态)和未来的特征(通过后向的状态)，通过时间的反向传播(BPTT)来训练BI-LSTM网络，具体结构如图2所示。

[0049] 随着时间推移，在展开的网络上进行的前向和后向传递同常规网络中的前向和后向传递方式类似，除了需要对所有的时间步骤展开隐藏状态，我们还需要在数据点的开始和结束时进行特殊的处理。在对整个句子进行前向扫描和后向扫描的时候仅仅需要在句子的开头将隐藏状态重置为0，通过做了批量的实现，这使得多个句子可以同时被处理。通过抽取一些基本特征以及电子病历中一些特有的特征，完成了隐私信息识别模型的构建。

[0050] 条件随机场(CRF)模型的关注点在句子级别上，而不是个别位置，CRF的输入和输出是直接相连的，这与LSTM和BI-LSTM网络刚好相反，是通过记忆细胞和循环组件连接在一起的。通过将LSTM网络和CRF网络整合成为LSTM-CRF模型，如图3所示。

[0051] 通过LSTM层，这个模型可以有效的利用过去的输入特征，通过CRF层，模型可以有效的利用句子级的标签信息。CRF层由连接连续输出层的线条表示。CRF层具有一个状态转移矩阵作为参数。利用这样的一个层，可以有效地利用过去和未来的标签来预测当前的标签，这类似于双向LSTM网络能够利用过去和未来的输入特征。

[0052] 再将一个双向LSTM网络和一个CRF网络合并成为一个BI-LSTM-CRF网络，网络结构如图4所示。除像LSTM-CRF模型那样能够利用过去的输入特征和句子级别的标签信息之外，BI-LSTM-CRF模型还能够利用未来的输入特征，这项额外的功能可以提高标注的准确性。

[0053] 采用Bi-LSTM双向长短期记忆网络将上述产生的向量作为输入，并产生对目标词的预测向量作为输出，迭代模块操作主要包含了向量层、前向长短期记忆网络层、后向长短期记忆网络层以及连接层，输出的向量将根据前向长短期记忆网络层的输出和后向长短期记忆网络层的输出而改变。给定训练集，前向LSTM考虑目标词前面的上下文信息，即从ω_1到ω_t的上下文信息，得到了目标词的一个向量c_t，具体计算如下公式所示：

[0054] $i_t=\delta(W_{wi}\omega_t+W_{hi}h_{t-1}+W_{ci}c_{t-1}+b_i)$ (1)

[0055] $$f_t=\delta(W_{\omega f}\omega_t+W_{hf}h_{t-1}+W_{cf}c_{t-1}+b_f) \quad (2)$$

[0056] $c_t=f_tc_{t-1}+\tanh(W_{\omega c}\omega_t+W_{hc}h_{t-1}+b_c)i_t$ (3)

[0057] 其中，式(1)中$W=\{\omega_1,\ldots\omega_t,\omega_{t+1}\ldots\omega_n\}$表示了词语序列，$\omega_t\in R^d$表示某句话中的第t词的向量表示，该词向量是d维词向量，n表示了该句话中词的个数，h_{t-1}表示Bi-LSTM中的记忆模块中之前隐藏向量，c_{t-1}表示记忆模块中之前原向量。同时将目标词经过后向LSTM计算，考虑到了目标词后面的上下文信息，即从ω_{t+1}到ω_n的上下文信息，得到了另一个向量o_t，具体计算如公式(4)所示：

[0058] $o_t=\delta(W_{\omega o}\omega_t+W_{ho}h_{t-1}+W_{co}c_t+b_o)$ (4)

[0059] 最终将上述同时产生的两个向量c_t和o_t输入连接层，使用双曲正切函数，得到该目标词的向量h_t，具体计算如公式(5)所示：

[0060] $h_t=o_t\tanh(c_t)$ (5)

[0061] (2)基于注意力机制的双向LSTM网络(Att-BiLSTM)

[0062] 在实体识别后,需要进行关系抽取以完成信息抽取的任务,但若将信息抽取分为两个子任务进行,则会导致子任务之间的误差传播,且在实现关系抽取时若采用有监督的方式训练模型,会出现数据标注耗时费力,占用大部分人力的情况。针对上述问题,本文拟采用基于弱监督学习的联合信息抽取,借鉴远程监督的方法和模式,预计构建一个能保证信息抽取效果,且尽量减少人工标注数据集,具有良好的提取速度和可扩展性的弱监督学习联合信息抽取。

[0063] 在双向LSTM网络上引入Attention的机制,构建Att-BiLSTM网络来处理文本分类的相关问题,解决CNN模型不适合学习长距离的语义信息的问题。在Att-BiLSTM网络中,主要由5个部分组成:

[0064] 输入层(Input layer):指的是输入的句子,对于中文,指的是对句子分好的词;

[0065] Embedding层:将句子中的每一个词映射成固定长度的向量;

[0066] LSTM层:利用双向的LSTM对embedding向量计算,实际上是双向LSTM通过对词向量的计算,从而得到更高级别的句子的向量;

[0067] Attention层:对双向LSTM的结果使用Attention加权;

[0068] 输出层(Output layer):输出层,输出具体的结果。

[0069] 在Bi-LSTM中我们会用最后一个时序的输出向量作为特征向量,然后进行softmax分类。Attention是先计算每个时序的权重,然后将所有时序的向量进行加权和作为特征向量,然后进行softmax分类。在实验中,加上Attention确实对结果有所提升,其模型结构如图5所示。

[0070] 其中,编码层采用了双向的RNN网络,最后隐层的输出是两个向量的拼接,表示为

[0071] $$h_t = \left[\overleftarrow{h_t}, \overrightarrow{h_t}\right] \qquad (6)$$

[0072] 而Attention层的输出为

[0073] $$\begin{cases} c_j = \sum_{j=1}^{T_x} a_{ij} h_j \\ a_{ij} = \dfrac{\exp\left(e_{ij}\right)}{\sum_{k=1}^{T_x} \exp\left(e_{ik}\right)} \\ e_{ij} = a\left(s_{i-1}, h_j\right) \end{cases} \qquad (7)$$

[0074] 在上述公式中h_j是编码层的隐层第j时刻的输出,s_{i-1}是解码层第i-1时刻隐层的输出。由此可以发现在计算c_i的模型实际上是一个线性模型,而且c_i事实上是编码层中各时刻隐含层输出的加权平均值。

[0075] 将编码层向量作为输入,产生序列标签作为输出,产生的最终预测向量h_t以及前向LSTM预测向量和词语所在的位置序号相乘进行更新并连接,最终经过双曲正切运算得到预测向量与其所在的位置向量相乘加上其偏差值,得到预测标签向量作为输出T_t。产生的语义向量,输入Softmax层,进行相似度计算,将其产生的实体标签概率加上TransE链接相似度计算概率值进行归一化,输出该实体标签的概率,具体计算如下:

[0076] $$y_t = W_y T_t(\min \sum_{i=0}^{\Delta} (h + r - t_i)) + b_y \quad (8)$$

[0077] $$p_t^i = \frac{\exp(y_t^i)}{\sum_{j=1}^{N_t} \exp(y_t^j)} \quad (9)$$

[0078] 其中W_y是Softmax层的矩阵，N_t表示了标签的数量，T_t表示预测标签向量，y_t表示实体关系标签概率，最终得到p_t^i表示了归一化后的标签概率。

[0079] (3)联合知识抽取网络

[0080] 语义距离的计算方法依赖联合向量的生成模型，可采用任何向量之间的距离计算，如欧几里德距离。通过知识表示学习的方法将原始数据中涉及到的实体和关系进行再训练，转换为嵌入向量形式，再训练模型采用了表示学习中的TransE模型，将训练集随机初始化为向量的形式作为输入，并产生训练集中实体集和预定义的关系集所对应的词向量作为输出。给定实体集、关系集和训练集，将通过训练集随机的替换头实体或者尾实体构建负样本，计算正确三元组实体和关系之间的距离，负样本中实体关系的距离，并调整两者之间的误差，将实体关系表示成符合现实关系的向量，TransE损失函数如下：

[0081] $$\min \sum_{(h,r,t)\in\Delta} \sum_{(h',r',t')\in\Delta'} [\gamma + f(h,r,t) - f(h',r',t')] \quad (10)$$

[0082] 式(10)中，TransE的损失函数分为超参数和正样本距离和负样本距离之差两部分的和，其中γ表示超参数，f(h,r,t)表示正样本的距离，f(h',r',t')表示负样本的距离，Δ表示正样本集合，Δ'表示负样本集合，[x]表示max(0,x)。

[0083] 知识图谱中的三元组和已经对齐的实体是用于学习联合知识的词向量，通过TransE和它的扩展方法PTransE获得两个知识库分别学到自己的知识向量，通过联合向量将这些知识向量合并到同一个语义空间，联合向量由已经对齐的实体获得。在合并的语义空间通过实体之间的语义距离实现实体之间的对齐，语义距离的计算方法依赖联合向量的生成模型，对于使用能量函数如公式11所示：

[0084] $$E(e_1,e_2) = \| e_1 - e_2 \|_{L1/L2}, \forall e_1 \in E_1, e_2 \in E_2 \quad (11)$$

[0085] 能量函数的值小于阈值，认为两个实体相似。使用新对齐得到的实体对更新联合向量和找到新的实体对，迭代学习联合向量和实体对齐采用了硬对齐和软对齐两种策略。

[0086] 通过弱监督学习的联合信息抽取算法，实现无子任务之间误差传播、无需大量时间和人力标注数据的从文本中抽取出有价值的三元组信息，解决现有的病历信息管理数据格式不统一、不规范的问题。通过TransE与表示学习相结合的算法训练知识库，生成联系更符合现实世界语义向量空间，实现基于知识图谱的病人病历相似度计算。

[0087] 实施例2：

[0088] 本发明实施例2提供了一种基于知识图谱的病人病历相似度评估系统，包括：

[0089] 数据获取模块，被配置为：获取病人病历文本数据并进行预处理；

[0090] 信息抽取模块，被配置为：采用基于弱监督学习的联合信息抽取模型对预处理后的数据进行实体识别和实体关系抽取，将得到的医学实体类型用知识向量进行表示；

[0091] 三元组构建模块，被配置为：根据得到的实体关系进行知识图谱中三元组的构建，并使用联合知识向量进行表示；

[0092] 相似度判别模块，被配置为：利用联合向量将知识向量合并到同一个语义空间，通过计算实体之间的语义距离进行相似度的判别。

[0093] 所述系统的工作方法与实施例1提供的基于知识图谱的病人病历相似度评估方法相同，这里不再赘述。

[0094] 实施例3：

[0095] 本发明实施例3提供了一种计算机可读存储介质，其上存储有程序，该程序被处理器执行时实现如本发明实施例1所述的基于知识图谱的病人病历相似度评估方法中的步骤，所述步骤为：

[0096] 获取病人病历文本数据并进行预处理；

[0097] 采用基于弱监督学习的联合信息抽取模型对预处理后的数据进行实体识别和实体关系抽取，将得到的医学实体类型用知识向量进行表示；

[0098] 根据得到的实体关系进行知识图谱中三元组的构建，并使用联合知识向量进行表示；

[0099] 利用联合向量将知识向量合并到同一个语义空间，通过计算疾病实体之间的语义距离进行病历相似度的判别。

[0100] 详细步骤与实施例1提供的基于知识图谱的病人病历相似度评估方法相同，这里不再赘述。

[0101] 实施例4：

[0102] 本发明实施例4提供了一种电子设备，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程序，所述处理器执行所述程序时实现如本发明第一方面所述的基于知识图谱的病人病历相似度评估方法中的步骤，所述步骤为：

[0103] 获取病人病历文本数据并进行预处理；

[0104] 采用基于弱监督学习的联合信息抽取模型对预处理后的数据进行实体识别和实体关系抽取，将得到的医学实体类型用知识向量进行表示；

[0105] 根据得到的实体关系进行知识图谱中三元组的构建，并使用联合知识向量进行表示；

[0106] 利用联合向量将知识向量合并到同一个语义空间，通过计算疾病实体之间的语义距离进行病历相似度的判别。

[0107] 详细步骤与实施例1提供的基于知识图谱的病人病历相似度评估方法相同，这里不再赘述。

[0108] 本领域内的技术人员应明白，本发明的实施例可提供为方法、系统、或计算机程序产品。因此，本发明可采用硬件实施例、软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器和光学存储器等)上实施的计算机程序产品的形式。

[0109] 本发明是参照根据本发明实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实

现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0110] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0111] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0112] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的程序可存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，所述的存储介质可为磁碟、光盘、只读存储记忆体(Read-Only Memory，ROM)或随机存储记忆体(Random AccessMemory，RAM)等。

[0113] 以上所述仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

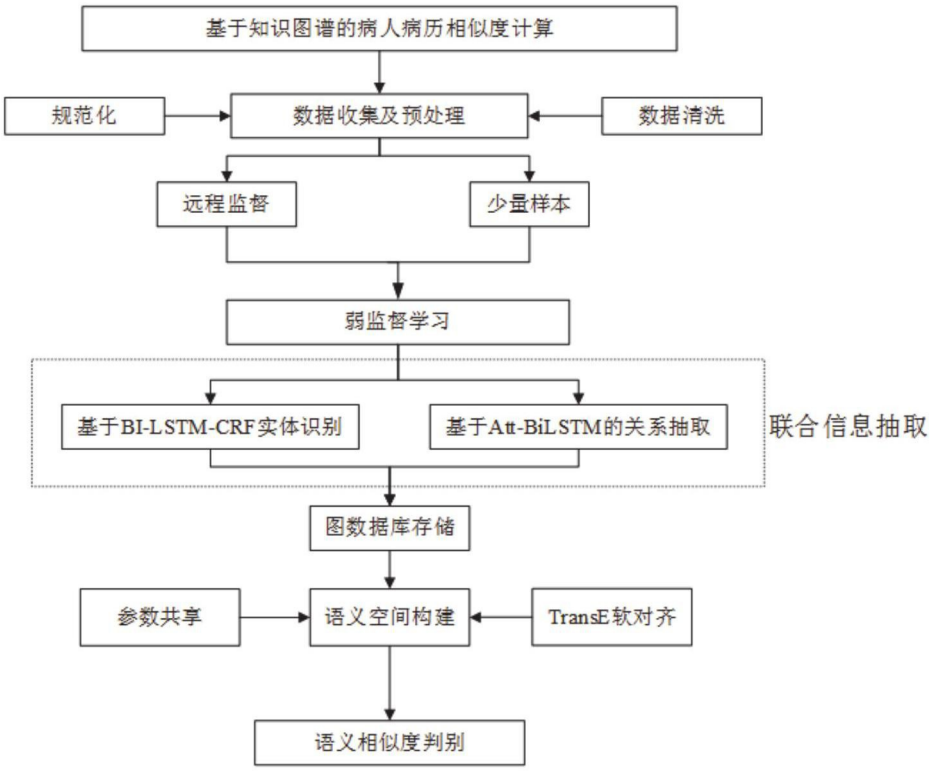

基于知识图谱的病人病历相似度计算
规范化
数据收集及预处理
数据清洗
远程监督
少量样本
弱监督学习
基于BI-LSTM-CRF实体识别
基于Att-BiLSTM的关系抽取
联合信息抽取
图数据库存储
参数共享
语义空间构建
TransE软对齐
语义相似度判别

图1

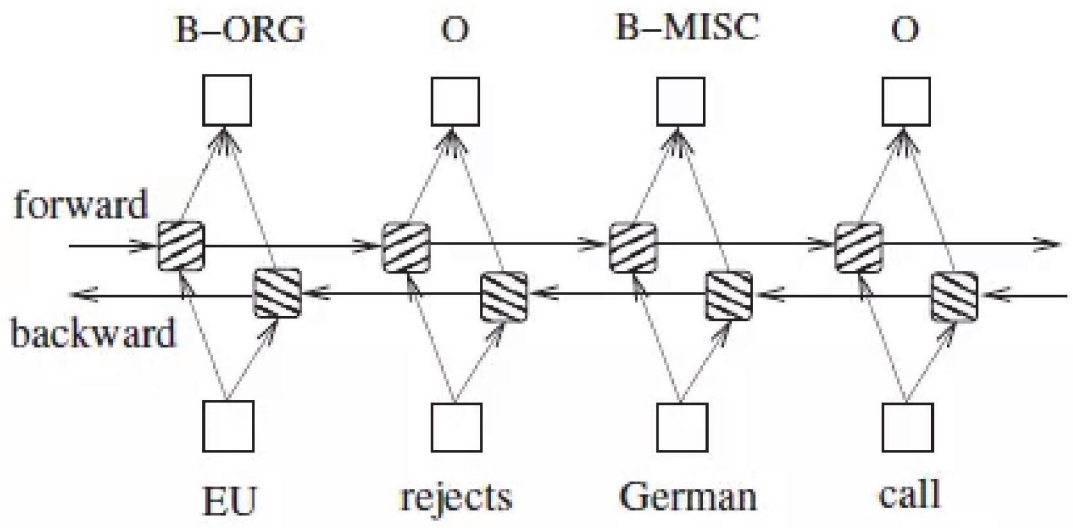

B-ORG
O
B-MISC
O
forward
backward
EU
rejects
German
call

图2

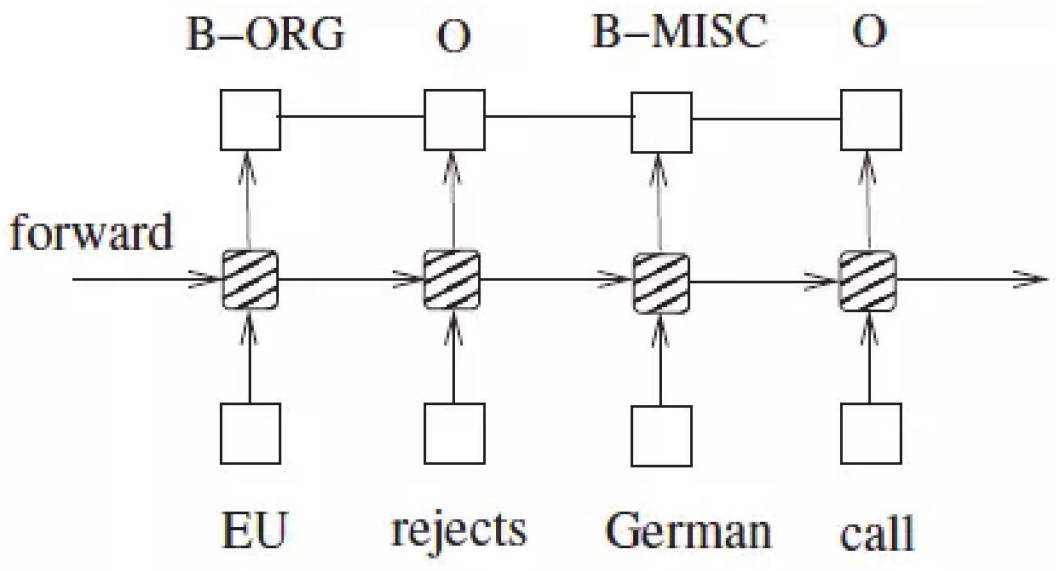

B-ORG
O
B-MISC
O
forward
EU
rejects
German
call

图3

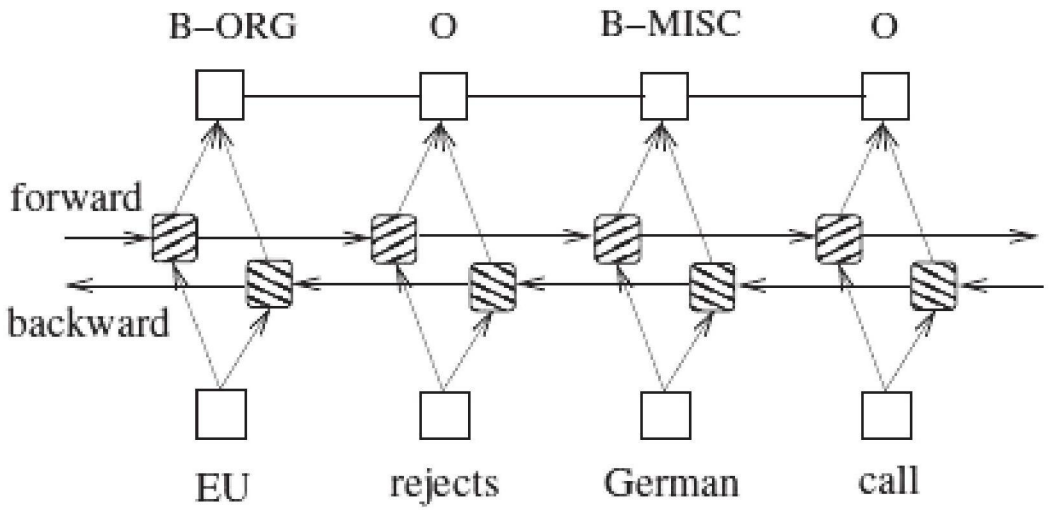

B-ORG
O
B-MISC
O
forward
backward
EU
rejects
German
call

图4

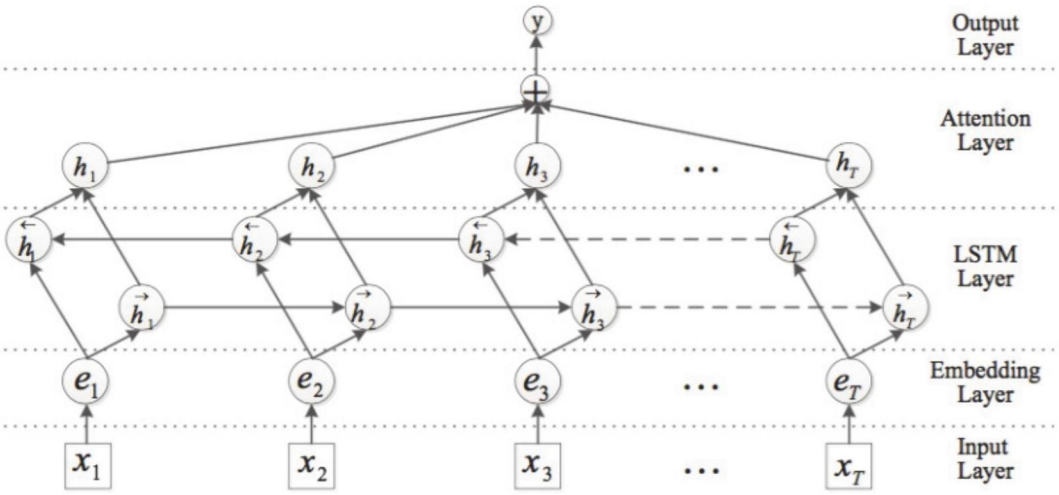

y
Output Layer
Attention Layer
LSTM Layer
Embedding Layer
Input Layer

图5