一种基于注意力机制神经网络与知识图谱的医疗文本分级方法

技术领域

本发明属于计算机人工智能领域，涉及一种基于注意力机制神经网络与知识图谱的医疗文本分级方法。

背景技术

近年来，随着人工智能技术在自然语言处理（Natural Language Processing, NLP）领域的广泛应用，人们对领域知识的探索也越来越多。其中医疗文本是在医疗领域中最常见的数据集，那么对医疗文本采用神经网络得到所对应的疾病分级结果，那么对于医学人工智能化具有非常重要的意义。

医疗文本分级可以认为是一个分类任务，通过抽取医疗文本中的特征数据在通过某种方法得到所对应疾病的严重程度分级结果。传统的文本分级的两种方法是基于传统机器学习方法和神经网络深度学习方法，目前深度学习方法占主动地位。基于深度学习的文本分级模型都是通过去文本数据抽取特征信息然后将其通过某个深度学习模型分类到具体疾病的分级结果。然后在当今的大数据时代，单纯的使用文本数据中的特征信息对于文本分级来说数据量太少，由于医疗文本的特殊性，并不能很好的得到全部的特征信息，那么对于医疗文本分级的准确性有很大的影响，如今的医疗文本分级仍然存在以下问题：（1）没有采用该疾病的知识图谱数据，单纯的使用文本数据导致数据量太少且准确性不高；（2）领域实体结构复杂并且相同概念存在很多中表达方法，尤其在医疗领域对于疾病病名与症状的描述例如“慢性阻塞性肺疾病”并不能很好的提取出来；（3）在目前神经网络中单纯的使用了实体信息，并没有利用在知识图谱中最重要的“关系”信息存在了语义上的割裂。基于以上问题，传统的医疗文本分级方法已经很难适用于如今的应用场景。

丁连红、[孙斌](http://xueshu.baidu.com/s?wd=author%3A%28%E5%AD%99%E6%96%8C%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson)、[张宏伟](http://xueshu.baidu.com/s?wd=author%3A%28%E5%BC%A0%E5%AE%8F%E4%BC%9F%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson)等人撰写的《[基于知识图谱扩展的短文本分类方法](http://www.cqvip.com/QK/72017X/201805/676674925.html)》中提到的基于学习知识图谱中的实体信息并提取文本数据中的实体数据作为特征输入在神经网络中进行分类的算法，使用了对应疾病的知识图谱作为外部数据源并改变了特征提取方法来实现医疗文本的分级任务。该方法虽然增加了只是领域实现了医疗文本的分级任务，但是该方法只是单纯的提取了实体信息，缺失了关键的关系特征，无法完全且准确的提取到文本中的所以信息。该方法虽还不够完善，但是使用了知识图谱的思路在文本分级的问题中给我们提供了思路，即需要使用外部知识来更好的提取文本特征。

发明内容

本发明的内容：

一种基于注意力机制神经网络与知识图谱的医疗文本分级方法，该方法包括：

1. 提出了一种基于注意力机制神经网络与知识图谱的医疗文本分级方法，该方法，该方法通过抽取知识图谱与文本数据中的实体-关系-三元组数据，通过增强型的LSTM模型得到对医疗文本的分级结果，不仅提高了分级结果的准确性，并且因为对于文本数据而言，只是单纯的提取了三元组数据无需对所有的数据进行提取好向量化还降低了计算成本
2. 首先得到知识图谱与文本数据中实体-关系-实体之间的直接关系数据，并且通过标准点互信息将知识图谱与文本数据中的直接关系合并为统一的三元组表达方式。
3. 其次通过TransE算法得到文本数据中的所有间接语义关系三元组，即通过在第2步中的直接关系通过推理机制得到所有的间接关系，得到文本数据中的所有的实体关系-实体三元组数据。
4. 采用增强型的LSTM模型通过基于Attention机制得到对于不同三元组的对于文本分级的重要性程度，将文本的三元组数据Encoder-Decoder模型得到文本的分级结果。

本发明的原理是一种基于注意力机制神经网络与知识图谱的医疗文本分级方法，不仅仅学习了基于知识图谱抽取的实体特征，也考虑了在知识图谱中实体之间的关系特征，并将直接关系特征通过基于规则的方式得到所有的间接语义关系特征，至此完成了对文本数据的特征抽取，采用该方法不仅仅降低了对于文本数据特征抽取的数量仅通过有限的实体-关系-实体三元组即可完成分级降低了运算量，又扩大了文本分级中的数据源且增加了专家知识提高了文本分级的准确性。

为达到以上发明目的，本发明采用如下的技术方案：

一种基于注意力机制神经网络与知识图谱的医疗文本分级方法，包括：直接关系抽取模块、直接关系统一模块、间接语义关系推理模块、神经网络分级模块。其中神经网络分级模块中使用本发明提出的基于Attention机制的增强型LSTM模型进行分级

直接关系抽取模块，在知识图谱中的实体关系包括两个部分：直接关系、间接语义关系，在该模块得到基于已有技术得到知识图谱与文本数据中的直接关系数据。

直接关系统一模块，由于医疗数据的特殊性，在很多医疗专有名词存在意思相同但表达相近、意思不同但表达相近的问题，所以在该模块采用基于点互信息的方法将知识图谱与文本数据中得到的直接关系统一起来得到统一的三元组数据。

间接语义关系推理模块，将统一直接关系模块中得到的所有的直接关系通过TransE算法得到文本数据中所有的间接语义关系三元组数据，得到文本数据的所有特征信息。

神经网络分级模块，在该模块中将文本数据中的所有实体-关系-实体三元组数据输入基于Attention机制的增强型LSTM模型得到不同的三元组数据对于文本分级的重要性程度，通过Encoder-Decoder模型得到医疗文本的分级结果。

本发明提出的通过神经网络学习知识图谱中的实体-关系-实体信息提出了一种新的文本分级方法，不仅能够学了知识图谱的专家知识提高了文本分级的准确性，而且降低了文本数据特征提取的数据量降低了计算量加快了运行速度。

附图说明

图1 基于注意力机制神经网络与知识图谱的医疗文本分级方法整体框架；

图2 基于Attention机制的增强型LSTM模型结构；

图3 增强型LSTM模型具体结构；

具体实施方式

下面将详细描述本发明各个方面的特征和示例性实施例：

本发明将文本中的直接关系与间接语义关系作为文本数据的特征，通过基于Attention机制的增强型LSTM模型得到文本分级结果，提高了文本分级准确率并降低了计算量。整体框架如图1所示分为：直接关系抽取模块、直接关系统一模块、间接语义关系推理模块、神经网络分级模块。具体的基于Attention的增强型LSTM模型结构如图2所示。具体的增强型LSTM模型如图3所示。TransE算法中直接关系与间接关系图4所示。

直接关系抽取模块（1）：从知识图谱中得到实体-关系-实体之间的直接关系三元组数据，从文本数据中基于规则得到实体-关系-实体之间的直接关系三元组数据。

直接关系统一模块（2）：该模块将上一模块中得到的所有直接关系三元组数据基于标准化的点互信息将从知识图谱与文本数据中的实体-关系-实体直接关系三元组统一起来，得到相同的实体与关系节点。

基于标准化点互信息的直接关系统一算法（21）：由于文本数据中的实体与关系的描述的不准确性和不唯一性，需要将从文本数据中抽取的实体之间直接关系三元组与从知识图谱中抽取的实体之间直接关系三元组合并为统一的直接关系三元组。

具体内容是将从文本数据中抽取出来的实体i与从知识图谱中抽取出来的实体c以此计算关联度，当关联度大于阈值时，我们认为i、c这两个实体之间是等价，可以将其合并为同一个实体来表示，建立统一的知识库。其中阈值的最佳取值经过交叉验证之后为0.85.

具体公式如下：

其中P(c|i) 表示对于一个实体 i它所对应的实体为c 的概率。P(i|c) 则表示对于一个实体c它所对应的实体为i 的概率。n(c,i) 表示实体 i 和实体 c同时出现的次数。SETC为全部实体c的集合，SETI为全部实体i的集合。表示实体 i 与集合SETC中每一个实体c共现次数之和。表示实体c与集合SETI中每一个实体i共现次数之和。

计算得到两个实体之间的条件概率， 用标准化的点互信息（NPMI）来计算两个实体i、c 之间的关联度评分，其具体公式如下

其中为实体i出现次数占所有实体出现次数的概率值；为实体i与实体c共同出现的次数占所有实体对出现次数的概率值；其中PMI为衡量两个实体之间相关性的点互信息，其具体公式如下：

间接语义关系推理模块（3）：该模块主要作用是将上一模块中得到的所有的直接关系三元组通过TransE算法推导，得到两个实体之间存在间接语义关系，从而得到文本数据中的所有数据。

TransE（Translation Embedding）是基于实体和关系的分布式向量表示，将三元组(head，relation，tail)看成向量h通过r翻译到t的过程，通过不断的调整向量h、r和t，找到一个使得从实体h 到实体t之间造成损失最小的关系r。通过定义势能函数使知识库中定义的势能大于不在知识库中的三元组的势能，即最小化整体势能，其具体公式如下：

其中为平滑系数在此设为1，为在知识库中的三元组数据，为不在知识库中的三元组数据，势能函数的计算方法为计算三元组数据线性组合的第二范式。

基于TransE算法计算两个实体之间的整体势能，得到最小的势能值所对应的关系，即是两个实体间的间接语义关系。

神经网络分级模块（4）：该模块应用基于Attention机制的增强型LSTM神经网络模型得到文本分级结果。附图2为具体的神经网络模型，通过增强型LSTM模型作为Encoder模型，基于Attention机制计算得到实体-关系-实体三元组对于文本分级的重要程度，并且通过普通的LSTM作为Decoder模型通过softmax变换得到文本分级结果。

附图3为具体的增强型LSTM模型，在增强型LSTM模型中细胞向量有两个，分别作为实体细胞向量与关系细胞向量，两个细胞向量分别计算实体与关系的长短信息并分别保存，在输出门采用将两个细胞向量通过一个全连接层得到最后的输出向量。

在t时刻的增强型LSTM模型的具体公式如下：

在公式（6）中计算为实体遗忘门系数，其中、分别为实体1与实体2在该公式中的权重矩阵，为t-1时刻输出向量在该公式中的权重矩阵，为在该公式中的偏置。

在公式（7）中计算为关系遗忘门系数，其中为关系在该公式中的权重矩阵，为t-1时刻输出向量在该公式中的权重矩阵，为在该公式中的偏置。

在公式（8）中计算为实体输入门系数，其中、分别为实体1与实体2在该公式中的权重矩阵，为t-1时刻输出向量在该公式中的权重矩阵，为在该公式中的偏置。

在公式（9）中计算为关系输入门系数，其中为关系在该公式中的权重矩阵，为t-1时刻输出向量在该公式中的权重矩阵，为在该公式中的偏置。

在公式（10）中计算为实体输入门备选状态，其中、分别为实体1与实体2在该公式中的权重矩阵，为t-1时刻输出向量在该公式中的权重矩阵，为在该公式中的偏置。

在公式（11）中计算为关系输入门备选状态，其中为关系在该公式中的权重矩阵，为t-1时刻输出向量在该公式中的权重矩阵，为在该公式中的偏置。

在公式（12）中计算为实体细胞状态，为t-1时刻的实体细胞状态，为遗忘门系数，为输入门系数，为实体输入门备选状态。

在公式（13）中计算为关系细胞状态，为t-1时刻的关系细胞状态，为遗忘门系数，为输入门系数，为关系输入门备选状态。

在公式（14）中计算为输出门系数，分别为实体1、实体2在该公式中的权重矩阵，为关系在该公式中的权重矩阵，为t-1时刻的输出向量在该公式中的权重矩阵，为该公式中的偏置。

在公式（15）中计算为输出门结果，为输出门系数，为实体细胞状态在该公式中的权重矩阵，为实体细胞状态，为关系细胞状态在该公式中的权重矩阵，为关系细胞状态。

在增强型LSTM模型，实体与关系向量采用word2vec初始化，其向量长度为200；所有的权重矩阵与偏置采用截断的正态分布做初始化，当参数收敛时或者当达到最大迭代次数10次时训练结束；模型中的为sigmoid函数，tanh为tanh函数。

采用Attention机制得到每个实体-关系-实体三元组对于文本分级的重要程度，在t时刻得到每一个三元组的重要性程度。记decoder模型时刻 t 的target hidden为，encoder的输出隐藏向量hidden state为encoder的全部输出向量集合为HS，对于其中任意,其权重的具体公式为：

其中：

为评分函数中的重要性系数，采用截断的正态分布做初始化；为在该公式中的权重矩阵，当参数收敛时或者当达到最大迭代次数10次时训练结束。

在Decoder层采用标准LSTM模型，将基于Attention机制之后的中间隐藏向量作为Decoder层的输入，在LSTM层之后的隐藏向量通过softmax得到最后的文本分级结果。