

基于 TextCNN 知识抽取方法、装置、计算机设备及存储介质

相关申请的交叉引用

- 5 本申请申明享有 2019 年 01 月 02 日递交的申请号为 CN 2019100026381、名称为“基于 TextCNN 知识抽取方法、装置、计算机设备及存储介质”的中国专利申请的优先权，该中国专利申请的整体现容以参考的方式结合在本申请中。

技术领域

- 10 本申请涉及知识图谱领域，尤其涉及一种知识抽取的方法、装置、计算机设备及存储介质。

背景技术

- 15 近年来，智能客服机器人在国内外的展应用近几年呈现出快速增长的势头，并逐步开始在电信运营商、金融服务等行业形成产业规模。智能客服对话不仅为企业与用户建立了快捷有效的沟通手段，还为企业提供了精细化管理所需的统计分析信息。

- 20 知识图谱是一种高效的信息存储与读取形式，能够有效应用于客服机器人场景。通过知识图谱，客服机器人能够根据聊天内容提供相应回应或者服务，提升用户体验，知识抽取，即从不同来源、不同结构的数据中进行知识提取，形成知识(结构化数据)存入到知识图谱，知识抽取是知识图谱建立过程中最关键且最主要的一步，但是，现有的知识抽取的建立过程较为繁琐，如果通过人工形式导入实体、关系信息，则需要耗费大量资源。

发明内客

- 25 本申请的目的是提供一种基于 TextCNN 知识抽取方法、装置、计算机设备及存储介质，用于解决现有技术存在的问题。

为实现上述目的，本申请提供一种基于 TextCNN 知识抽取方法，包括以下步骤：

S10 收集第一训练数据，构建字向量字典与词向量字典；

S20 构建第一卷积神经网络，并基于第一优化算法训练所述第一卷积神经网络；所述第一卷积神经网络包括依次连接的第一嵌入层、第一多层卷积以及第一 softmax 函数；

- 30 S21 收集第二训练数据，所述第二训练数据为预先标注数据，包括命名实体位置标签与命名实体关系标签，将所述第二训练数据分割为单一文字并去除特殊符号后，输入至所述第一嵌入层中；

S22 基于所述字向量字典，于所述第一嵌入层将字符级的第二训练数据进行字向量匹配，以将所述第二训练数据转化为矩阵形式；

- 35 S23 所述第一多层卷积用于对所述第一嵌入层输出的矩阵进行卷积运算，所述第一多层卷积包括位于前部的一组第一类一维卷积层以及位于后部的至少一组第二类一维卷积层，所述第一类一维卷积层包括长度数量不同、通道数量相同设置一维卷积核，所述第二类一维卷积层包括长度数量与通道数量均相同设置的一维卷积核，每一卷积层的数据均来自于前一卷积层输出，且所述卷积运算过程中，保持矩阵行数不变；

- 40 S24 将所述第一多层卷积经由第一 softmax 函数进行输出，以确定每一文字的多个细分类别 BEMO（其中，B 实体开端，M 实体中部，E 实体结尾，O 非实体）标注的预测概率；

S25 训练所述第一卷积神经网络，根据所述 BEMO 标注预测概率与所述第二训练数据真实 BEMO 标签，计算损失函数交叉熵，并通过第一优化算法对所述损失函数进行最小化，以训练所述第一卷积神经网络；

- 45 S30 构建第二卷积神经网络，并基于第二优化算法训练所述第二卷积神经网络，所述第二卷积神经网络包括依次连接的第二嵌入层、第二多层卷积、一池化层、两层全连接层

以及第二 softmax 函数；

S31 对所述第二训练数据进行分词；

5 S311 采用 jieba 库对所述第二训练数据进行初步分词，并比照第一卷积神经网络的预测进行修正，若所述初步分词结果与所述第一卷积神经网络的预测的分词结果不同，以所述第一卷积神经网络的预测的分词结果为准；

S312 去除所述初步分词中的特殊符号及非中文字符后，将所述处理后的第二训练数据输入至所述第二嵌入层；

S32 基于所述词向量词典，于所述第二嵌入层对所述分词后的第二训练数据进行词向量匹配，以将所述第二训练数据转化为矩阵形式；

10 S33 基于所述第二多层卷积对所述第二嵌入层输出的矩阵进行卷积运算，所述第二多层卷积包括位于前部的一组第一类一维卷积层以及位于后部的至少一组第二类一维卷积层，所述第一类一维卷积层包括长度数量不同、通道数量相同设置一维卷积核，所述第二类一维卷积层包括长度数量与通道数量均相同设置的一维卷积核，每一卷积层的数据均来自于前一卷积层输出，且所述卷积运算过程中保持矩阵行数不变；

15 S34 将所述第二多层卷积的输出输入至所述池化层进行压缩；

S35 将所述池化层的输出输入至两层所述全连接层中进行各个通道的信息融合；

S36 将所述全连接层的输出输入至所述第二 softmax 函数中，用于确定多个实体关系标签的对应预测概率，

20 S37 训练所述第二卷积神经网络，依据所述第二卷积神经网络输出的关系标签的预测概率与所述第二训练数据真实关系标签，计算第二损失函数交叉熵，并通过优化算法对所述损失函数进行最小化，以训练所述第二卷积神经网络；

25 S40 将待预测数据输入至训练完毕的所述第一卷积神经网络与所述第二卷积神经网络中，依据训练后所述第一卷积神经网络输出的实体标注预测与训练后所述第二卷积神经网络输出实体关系预测，提取待预测数据的知识图谱三元组：选取所述 BEMO 标注预测概率中最高概率值对应的类作为所述第一卷积神经网络输出的实体标注预测，选取预测概率值大于 0.5 的类作为所述第二卷积神经网络输出实体关系预测，以提取将待预测数据的知识图谱三元组。

本申请还提供了一种基于 TextCNN 的知识抽取装置，包括：

30 字向量字典构建模块：用于基于收集的第一训练数据，构建字向量字典；

词向量字典构建模块：用于基于收集的第一训练数据，构建词向量字典；

第一卷积神经网络构建及训练模块，用于构建第一卷积神经网络并基于第一优化算法训练所述第一卷积神经网络；所述第一卷积神经网络包括依次连接的第一嵌入层、第一多层卷积以及第一 softmax 函数；包括

35 字向量预处理单元，所述第二训练数据为预先标注数据，包括命名实体位置标签与命名实体关系标签，所述字向量预处理模块用于将所述第二训练数据分割为单一文字并去除特殊符号后，输入至所述第一嵌入层中；

字向量矩阵化单元，用于基于所述字向量字典，于所述第一嵌入层将字符级的第二训练数据进行字向量匹配，以将所述第二训练数据转化为矩阵形式；

40 第一多层卷积单元，用于对所述第一嵌入层输出的矩阵进行卷积运算，所述第一多层卷积单元包括位于前部的一组第一类一维卷积层以及位于后部的至少一组第二类一维卷积层，所述第一类一维卷积层包括长度数量不同、通道数量相同设置一维卷积核，所述第二类一维卷积层包括长度数量与通道数量均相同设置的一维卷积核，每一卷积层的数据均来自于前一卷积层输出，且所述卷积运算过程中，保持矩阵行数不变；

45 第一 softmax 函数输出单元，用于将所述第一多层卷积经由第一 softmax 函数进行输出，以确定每一文字的多个细分类别 BEMO 标注的预测概率；

第一卷积神经网络训练单元，用于根据所述 BEMO 标注预测概率与所述第二训练数据真实 BEMO 标签，计算损失函数交叉熵，并通过第一优化算法对所述损失函数进行最小化，以训练所述第一卷积神经网络；

5 第二卷积神经网络构建及训练模块，用于构建第二卷积神经网络并基于第二优化算法训练所述第二卷积神经网络，所述第二卷积神经网络包括依次连接的第二嵌入层、第二多层卷积单元、一池化层、两层全连接层以及第二 softmax 函数，包括：

词向量预处理单元：用于对所述第二训练数据进行分词，包括：

初步分词子单元，用于采用 jieba 库对所述第二训练数据进行初步分词，并比照第一卷积神经网络的预测进行修正，若所述初步分词结果与所述第一卷积神经网络的预测的分词结果不同，以所述第一卷积神经网络的预测的分词结果为准；

10 分词预处理子单元，用于去除所述初步分词中的特殊符号及非中文字符后，将所述处理后的第二训练数据输入至所述第二嵌入层；

字向量矩阵化单元，用于基于所述词向量词典，于所述第二嵌入层对所述分词后的第二训练数据进行词向量匹配，以将所述第二训练数据转化为矩阵形式；

15 第二多层卷积单元，用于基于所述第二多层卷积单元对所述第二嵌入层输出的矩阵进行卷积运算，所述第二多层卷积单元包括位于前部的一组第一类一维卷积层以及位于后部的至少一组第二类一维卷积层，所述第一类一维卷积层包括长度数量不同、通道数量相同设置一维卷积核，所述第二类一维卷积层包括长度数量与通道数量均相同设置的一维卷积核，每一卷积层的数据均来自于前一卷积层输出，且所述卷积运算过程中保持矩阵行数不变；

20 池化层，用于将所述第二多层卷积单元的输出输入至所述池化层进行压缩；

全连接层，用于将所述池化层的输出输入至两层所述全连接层中进行各个通道的信息融合；

25 第二 softmax 函数输出单元，用于将所述全连接层的输出输入至所述第二 softmax 函数中，确定多个实体关系标签的对应预测概率，

训练第二卷积神经网络单元，用于依据所述第二卷积神经网络输出的关系标签的预测概率与所述第二训练数据真实关系标签，计算第二损失函数交叉熵，并通过优化算法对所述损失函数进行最小化，以训练所述第二卷积神经网络；

30 知识图谱三元组提取模块，用于将待预测数据输入至训练完毕的所述第一卷积神经网络与所述第二卷积神经网络中，依据训练后所述第一卷积神经网络输出的实体标注预测与训练后所述第二卷积神经网络输出实体关系预测，提取待预测数据的知识图谱三元组：选取所述 BEMO 标注预测概率中最高概率值对应的类作为所述第一卷积神经网络输出的实体标注预测，选取预测概率值大于 0.5 的类作为所述第二卷积神经网络输出实体关系预测，以提取待预测数据的知识图谱三元组。

35 为实现上述目的，本申请还提供一种计算机设备，包括存储器、处理器以及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现基于 TextCNN 知识抽取方法，包括以下步骤：

S10 收集第一训练数据，构建字向量字典与词向量字典；

40 S20 构建第一卷积神经网络，并基于第一优化算法训练所述第一卷积神经网络；所述第一卷积神经网络包括依次连接的第一嵌入层、第一多层卷积以及第一 softmax 函数；

S21 收集第二训练数据，所述第二训练数据为预先标注数据，包括命名实体位置标签与命名实体关系标签，将所述第二训练数据分割为单一文字并去除特殊符号后，输入至所述第一嵌入层中；

45 S22 基于所述字向量字典，于所述第一嵌入层将字符级的第二训练数据进行字向量匹配，以将所述第二训练数据转化为矩阵形式；

S23 所述第一多层卷积用于对所述第一嵌入层输出的矩阵进行卷积运算，所述第一多

层卷积包括位于前部的一组第一类一维卷积层以及位于后部的至少一组第二类一维卷积层，所述第一类一维卷积层包括长度数量不同、通道数量相同设置一维卷积核，所述第二类一维卷积层包括长度数量与通道数量均相同设置的一维卷积核，每一卷积层的数据均来自于前一卷积层输出，且所述卷积运算过程中，保持矩阵行数不变；

5 S24 将所述第一多层卷积经由第一 softmax 函数进行输出，以确定每一文字的多个细分类别 BEMO（其中，B 实体开端，M 实体中部，E 实体结尾，O 非实体）标注的预测概率；

S25 训练所述第一卷积神经网络，根据所述 BEMO 标注预测概率与所述第二训练数据真实 BEMO 标签，计算损失函数交叉熵，并通过第一优化算法对所述损失函数进行最小化，以训练所述第一卷积神经网络；

10 S30 构建第二卷积神经网络，并基于第二优化算法训练所述第二卷积神经网络，所述第二卷积神经网络包括依次连接的第二嵌入层、第二多层卷积、一池化层、两层全连接层以及第二 softmax 函数；

S31 对所述第二训练数据进行分词；

15 S311 采用 jieba 库对所述第二训练数据进行初步分词，并比照第一卷积神经网络的预测进行修正，若所述初步分词结果与所述第一卷积神经网络的预测的分词结果不同，以所述第一卷积神经网络的预测的分词结果为准；

S312 去除所述初步分词中的特殊符号及非中文字符后，将所述处理后的第二训练数据输入至所述第二嵌入层；

20 S32 基于所述词向量词典，于所述第二嵌入层对所述分词后的第二训练数据进行词向量匹配，以将所述第二训练数据转化为矩阵形式；

S33 基于所述第二多层卷积对所述第二嵌入层输出的矩阵进行卷积运算，所述第二多层卷积包括位于前部的一组第一类一维卷积层以及位于后部的至少一组第二类一维卷积层，所述第一类一维卷积层包括长度数量不同、通道数量相同设置一维卷积核，所述第二类一维卷积层包括长度数量与通道数量均相同设置的一维卷积核，每一卷积层的数据均来自于前一卷积层输出，且所述卷积运算过程中保持矩阵行数不变；

S34 将所述第二多层卷积的输出输入至所述池化层进行压缩；

S35 将所述池化层的输出输入至两层所述全连接层中进行各个通道的信息融合；

S36 将所述全连接层的输出输入至所述第二 softmax 函数中，用于确定多个实体关系标签的对应预测概率；

30 S37 训练所述第二卷积神经网络，依据所述第二卷积神经网络输出的关系标签的预测概率与所述第二训练数据真实关系标签，计算第二损失函数交叉熵，并通过优化算法对所述损失函数进行最小化，以训练所述第二卷积神经网络；

35 S40 将待预测数据输入至训练完毕的所述第一卷积神经网络与所述第二卷积神经网络中，依据训练后所述第一卷积神经网络输出的实体标注预测与训练后所述第二卷积神经网络输出实体关系预测，提取待预测数据的知识图谱三元组：选取所述 BEMO 标注预测概率中最高概率值对应的类作为所述第一卷积神经网络输出的实体标注预测，选取预测概率值大于 0.5 的类作为所述第二卷积神经网络输出实体关系预测，以提取将待预测数据的知识图谱三元组。

40 为实现上述目的，本申请还提供计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现基于 TextCNN 知识抽取方法，包括以下步骤：

S10 收集第一训练数据，构建字向量字典与词向量字典；

S20 构建第一卷积神经网络，并基于第一优化算法训练所述第一卷积神经网络；所述第一卷积神经网络包括依次连接的第一嵌入层、第一多层卷积以及第一 softmax 函数；

45 S21 收集第二训练数据，所述第二训练数据为预先标注数据，包括命名实体位置标签与命名实体关系标签，将所述第二训练数据分割为单一文字并去除特殊符号后，输入至所述第一嵌入层中；

S22 基于所述字向量字典，于所述第一嵌入层将字符级的第二训练数据进行字向量匹配，以将所述第二训练数据转化为矩阵形式；

5 S23 所述第一多层卷积用于对所述第一嵌入层输出的矩阵进行卷积运算，所述第一多层卷积包括位于前部的一组第一类一维卷积层以及位于后部的至少一组第二类一维卷积层，所述第一类一维卷积层包括长度数量不同、通道数量相同设置一维卷积核，所述第二类一维卷积层包括长度数量与通道数量均相同设置的一维卷积核，每一卷积层的数据均来自于前一卷积层输出，且所述卷积运算过程中，保持矩阵行数不变；

S24 将所述第一多层卷积经由第一 softmax 函数进行输出，以确定每一文字的多个细分类别 BEMO（其中，B 实体开端，M 实体中部，E 实体结尾，O 非实体）标注的预测概率；

10 S25 训练所述第一卷积神经网络，根据所述 BEMO 标注预测概率与所述第二训练数据真实 BEMO 标签，计算损失函数交叉熵，并通过第一优化算法对所述损失函数进行最小化，以训练所述第一卷积神经网络；

S30 构建第二卷积神经网络，并基于第二优化算法训练所述第二卷积神经网络，所述第二卷积神经网络包括依次连接的第二嵌入层、第二多层卷积、一池化层、两层全连接层以及第二 softmax 函数；

S31 对所述第二训练数据进行分词；

S311 采用 jieba 库对所述第二训练数据进行初步分词，并比照第一卷积神经网络的预测进行修正，若所述初步分词结果与所述第一卷积神经网络的预测的分词结果不同，以所述第一卷积神经网络的预测的分词结果为准；

20 S312 去除所述初步分词中的特殊符号及非中文字符后，将所述处理后的第二训练数据输入至所述第二嵌入层；

S32 基于所述词向量词典，于所述第二嵌入层对所述分词后的第二训练数据进行词向量匹配，以将所述第二训练数据转化为矩阵形式；

25 S33 基于所述第二多层卷积对所述第二嵌入层输出的矩阵进行卷积运算，所述第二多层卷积包括位于前部的一组第一类一维卷积层以及位于后部的至少一组第二类一维卷积层，所述第一类一维卷积层包括长度数量不同、通道数量相同设置一维卷积核，所述第二类一维卷积层包括长度数量与通道数量均相同设置的一维卷积核，每一卷积层的数据均来自于前一卷积层输出，且所述卷积运算过程中保持矩阵行数不变；

S34 将所述第二多层卷积的输出输入至所述池化层进行压缩；

30 S35 将所述池化层的输出输入至两层所述全连接层中进行各个通道的信息融合；

S36 将所述全连接层的输出输入至所述第二 softmax 函数中，用于确定多个实体关系标签的对应预测概率，

35 S37 训练所述第二卷积神经网络，依据所述第二卷积神经网络输出的关系标签的预测概率与所述第二训练数据真实关系标签，计算第二损失函数交叉熵，并通过优化算法对所述损失函数进行最小化，以训练所述第二卷积神经网络；

S40 将待预测数据输入至训练完毕的所述第一卷积神经网络与所述第二卷积神经网络中，依据训练后所述第一卷积神经网络输出的实体标注预测与训练后所述第二卷积神经网络输出实体关系预测，提取待预测数据的知识图谱三元组：选取所述 BEMO 标注预测概率中最高概率值对应的类作为所述第一卷积神经网络输出的实体标注预测，选取预测概率值大于 0.5 的类作为所述第二卷积神经网络输出实体关系预测，以提取将待预测数据的知识图谱三元组。

45 本申请提供一种基于 TextCNN 知识抽取方法、装置、计算机设备及存储介质，通过卷积神经网络实现知识图谱中的知识抽取环节，在保证精度的前提下有效提升了模型训练效率。通过将训练文本转化为向量形式，并接入两类卷积神经网络模型（卷积层形式均为一维卷积核）中以提炼训练文本信息，分别实现命名实体识别以及实体关系识别。由于卷积神经网络具有并行计算特性，因此能够充分利用计算资源实现计算效率提升，训练完成

的两类卷积神经网络模型通过预测融合实现自动化知识抽取。

具体而言，通过将待预测数据分别转化为字向量与词向量形式，并接入至第一卷积神经网络与第二卷积神经网络处理，其中第一卷积神经网络用以实现命名实体识别。所用卷积神经网络为全卷积形式，其输入为字向量，输出为实体类别边界预测。经过这一过程可将原本连续的文本进行分词，并保留命名实体相关文字并进行实体归类；第二卷积神经网络实现知识抽取。所用卷积神经网络包括卷积层、池化层等，其输入包括字向量与词向量，输出为关系抽取识别。经过这一过程，可确定文本中知识实体的关联。结合待预测数据的实体标注预测与实体关系预测，即可识别出待预测数据的实体标注预测与实体关系预测中存在的实体及其相互关系，用于提取待预测数据的知识图谱三元组，从而实现自动化的知识抽取，且在保证精度的前提下有效提升了模型训练效率。

附图说明

图 1 为本申请基于 TextCNN 知识抽取方法一实施例的流程图；

图 2 为本申请基于 TextCNN 知识抽取装置一实施例的程序模块示意图；

图 3 为本申请基于 TextCNN 知识抽取装置一实施例的硬件结构示意图。

具体实施方式

为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本申请，并不用于限定本申请。基于本申请中的实施例，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他实施例，都属于本申请保护的范围。

实施例一

本申请提供一种基于 TextCNN 的知识抽取方法，如图 1 所示，包括以下步骤：

S10 收集第一训练数据，构建字向量字典与词向量字典；

优选的，步骤 S10 中，包括：

S11 将收集的第一训练数据划分为单独文字并去除特殊符号与非中文字符后，代入 Word2Vec 算法进行训练，获取字向量并构建字向量字典；

S12 同时，将收集的第一训练数据进行分词并去除特殊符号与非中文字符后，代入 Word2Vec 算法进行训练，获取词向量字典。本实施例中，Word2Vec 算法训练通过 Python 中的 gensim 库实现。

本申请所示基于 TextCNN 的知识抽取方法的首先获取字、词向量字典，即确定字词与向量的对应关系。字、词向量分别单独构建向量，第一训练文本均为中文维基百科。对于字向量，首先将训练文本划分为单独文字并去除特殊符号与非中文字符，随后将处理后的文本代入 Word2Vec 算法进行训练，获取字向量。而对于词向量，首先将训练文本进行分词并去除特殊符号与非中文字符，随后将处理后的文本代入 Word2Vec 算法进行训练，获取词向量。字向量与词向量的维度均为 300。这一步骤中涉及的分词通过 Python 中的 jieba 库实现，而这一步骤中涉及的 Word2Vec 算法训练则通过 Python 中的 gensim 库实现。

S20 构建第一卷积神经网络，并基于第一优化算法训练第一卷积神经网络；第一卷积神经网络包括依次连接的第一嵌入层、第一多层卷积以及第一 softmax 函数；本实施例中，可基于 Python 中的 tensorflow 库建立第一卷积神经网络。

S21 收集第二训练数据，第二训练数据为预先标注数据，包括命名实体位置标签与命名实体关系标签，将第二训练数据分割为单一文字并去除特殊符号后，输入至第一嵌入层中；

S22 基于字向量字典，于第一嵌入层将字符级的第二训练数据进行字向量匹配，以将第二训练数据转化为矩阵形式；

S23 第一多层卷积用于对第一嵌入层输出的矩阵进行卷积运算，第一多层卷积包括位

于前部的一组第一类一维卷积层以及位于后部的至少一组第二类一维卷积层，第一类一维卷积层包括长度数量不同、通道数量相同设置一维卷积核，第二类一维卷积层包括长度数量与通道数量均相同设置的一维卷积核，每一卷积层的数据均来自于前一卷积层输出，且卷积运算过程中，保持矩阵行数不变；

- 5 本实施例中，作为一优选方案，第一多层卷积包括 5 层卷积层，第一类一维卷积层包括了 3 类长度的一维卷积核，各对应 128 通道；四组的第二类一维卷积层的一维卷积核长度均为 3 且通道数为 384；

S24 将第一多层卷积经由第一 softmax 函数进行输出，以确定每一文字的多个细分类别 BEMO 标注的预测概率；

- 10 S25 训练第一卷积神经网络，根据 BEMO 标注预测概率与第二训练数据真实 BEMO 标签，计算损失函数交叉熵，并通过第一优化算法对损失函数进行最小化，以训练第一卷积神经网络；第一优化算法为 ADAM 算法或为 rmsprop 算法。

- 步骤 S20 中，第二训练文本与字、词向量的第一训练文本不同，其形式为短句形式并包含了命名实体位置标签（对于每一个文字进行 BEMO 中文命名实体边界标注）以及命名实体关系标签。预处理方面，本步骤中将第二训练文本分割为单一文字并去除特殊符号，
15 并将处理后的文本输入第一卷积神经网络。第一卷积神经网络于嵌入层将字符级的第二训练文本进行字向量匹配，从而将第二训练文本转化为矩阵形式（矩阵每一行对应一个文字的向量）。字向量匹配完毕后，第一卷积神经网络即可进行卷积运算，本实施例中，第一多层卷积共设计了 5 层卷积层，每一卷积层的数据均来自于前一卷积层输出。其中，位于
20 第一层的第一类一维卷积层包括了 3 类长度（1、3、5）的一维卷积核各对应 128 通道，而其余第一类一维卷积层的一维卷积核长度均为 3 且通道数为 384。需要注意的是，卷积层运算过程中需要保持矩阵行数不变。第一多层卷积的最后的卷积层通过第一 softmax 函数进行输出，这一输出对应每一文字的 BEMO 标注预测概率。BEMO 标注包括多类细分，因此第一卷积神经网络将计算某文字对于每一细分类别的概率，例如“B_地名”的概率、“B_人名”的概率，“E_人名”的概率等。模型构建完毕后，即可进行训练。训练过程中，算法将根据第一卷积神经网络的 BEMO 标注预测概率与第二训练文本的真实 BEMO 标签计算损失函数交叉熵，并通过优化算法 ADAM 对损失函数进行最小化，训练神经网络。需要注意的是，这一模型可能存在文字标注预测矛盾的现象，因此模型仅提取前后文字标注对应的实体。第一卷积神经网络的建立通过 Python 中的 tensorflow 库实现。至此，方案完成了第一卷积神经网络的构建及训练。
30

S30 构建第二卷积神经网络，并基于第二优化算法训练第二卷积神经网络，第二卷积神经网络包括依次连接的第二嵌入层、第二多层卷积、一池化层、两层全连接层以及第二 softmax 函数；本实施例中，可基于 Python 中的 tensorflow 库建立第二卷积神经网络。

S31 对第二训练数据进行分词；

- 35 S311 采用 jieba 库对第二训练数据进行初步分词，并比照第一卷积神经网络的预测进行修正，若初步分词结果与第一卷积神经网络的预测的分词结果不同，以第一卷积神经网络的预测的分词结果为准；

S312 去除初步分词中的特殊符号及非中文字符后，将处理后的第二训练数据输入至第二嵌入层；

- 40 S32 基于词向量词典，于第二嵌入层对分词后的第二训练数据进行词向量匹配，以将第二训练数据转化为矩阵形式；

- S33 基于第二多层卷积对第二嵌入层输出的矩阵进行卷积运算，第二多层卷积包括位于前部的一组第一类一维卷积层以及位于后部的至少一组第二类一维卷积层，第一类一维卷积层包括长度数量不同、通道数量相同设置一维卷积核，第二类一维卷积层包括长度数量与通道数量均相同设置的一维卷积核，每一卷积层的数据均来自于前一卷积层输出，且
45 卷积运算过程中保持矩阵行数不变；本实施例中，作为一优选方案，第二多层卷积包括 3

层卷积层，第一类一维卷积层包括了3类长度的一维卷积核，各对应128通道；两组的第二类一维卷积层的一维卷积核长度均为3且通道数为384。

S34 将第二多层卷积的输出输入至池化层进行压缩；

S35 将池化层的输出输入至两层全连接层中进行各个通道的信息融合；

5 S36 将全连接层的输出输入至第二 softmax 函数中，用于确定多个实体关系标签的对应预测概率，

S37 训练第二卷积神经网络，依据第二卷积神经网络输出的关系标签的预测概率与第二训练数据真实关系标签，计算第二损失函数交叉熵，并通过优化算法对损失函数进行最小化，以训练第二卷积神经网络；本实施例中，第二优化算法为 ADAM 算法或为 rmsprop
10 算法。

本实施例中，第二卷积神经网络构建所采取的训练文本与第二卷积神经网络相同，但是所采用的标注不是 BMEIO 标注，而是命名实体关系标注。数据预处理方面，由于第一卷积神经网络已标注了命名实体边界，但未标注非命名实体词语边界，因此方案采用 jieba 库对文本进行初步分词并根据第一卷积神经网络的识别结果对其进行修正，最后去除特殊符号及非中文字符。文本处理完毕后即可输入第二卷积神经网络，并于嵌入层对文本进行词
15 向量匹配（词向量字典中不存在的命名实体向量初始化为 0）。第二多层卷积包含前部 3 层卷积层。其中，位于第一层的第一类一维卷积层包括了3类长度（1、3、5）的一维卷积核各对应128通道，而其余卷积层的一维卷积核长度均为3且通道数为384。需要注意的是，卷积层运算过程中需保持矩阵行数不变。卷积运算输出将输入池化层进行压缩，池化形式为 Max-Pooling，池化层的输出将输入两层全连接层，实现各个通道的信息融合。全连接层通过第二 softmax 函数形式输出多标签的对应预测概率，例如“地理关系”概率、“从属关系”概率等多类标签。模型构建完毕后即可进行训练。训练过程中，算法将根据模型的关系预测与真实关系标签计算损失函数交叉熵，并通过优化算法 ADAM 对损失函数进行最小化，训练神经网络。第二卷积神经网络的建立通过 Python 中的 tensorflow 库实现。至此方
20 案完成了第二卷积神经网络与训练工作。

S40 将待预测数据输入至训练完毕的第一卷积神经网络与第二卷积神经网络中，依据训练后第一卷积神经网络输出的实体标注预测与训练后第二卷积神经网络输出实体关系预测，提取待预测数据的知识图谱三元组：选取 BEMO 标注预测概率中最高概率值对应的类作为第一卷积神经网络输出的实体标注预测，选取预测概率值大于 0.5 的类作为第二卷积
30 神经网络输出实体关系预测，以提取待预测数据的知识图谱三元组。此外，步骤 S40 中，若实体标注预测与实体关系预测结果相互矛盾，则放弃待预测数据的知识图谱三元组提取。

本步骤中，将两类卷积神经网络的预测结果进行融合。由于两类卷积神经网络预测结果可能存在矛盾，例如对于某语句，第一卷积神经网络预测包括人物实体，但是第二卷积神经网络预测属于“地理关系”。因此，方案仅将两类模型预测结果对应的知识进行提取。
35 例如对于某语句，第一卷积神经网络预测包括人物实体，且第二卷积神经网络预测属于“从属关系”，则对于这一知识图谱三元组进行提取。

本申请所示的本申请提供的一种基于 TextCNN 知识抽取方法通过卷积神经网络实现知识图谱中的知识抽取环节，在保证精度的前提下有效提升了模型训练效率。通过将训练文本转化为向量形式，并接入两类卷积神经网络模型（卷积层形式均为一维卷积核）中以提炼训练文本信息，分别实现命名实体识别以及实体关系识别。由于卷积神经网络具有并行计算特性，因此能够充分利用计算资源实现计算效率提升，训练完成的两类卷积神经网络模型通过预测融合实现自动化知识抽取。具体而言，通过将待预测数据分别转化为字向量与词向量形式，并接入至第一卷积神经网络与第二卷积神经网络处理，其中第一卷积神经网络用以实现命名实体识别。所用卷积神经网络为全卷积形式，其输入为字向量，输出为
40 实体类别边界预测。经过这一过程可将原本连续的文本进行分词，并保留命名实体相关文字并进行实体归类；第二卷积神经网络实现知识抽取。所用卷积神经网络包括卷积层、池

化层等，其输入包括字向量与词向量，输出为关系抽取识别。经过这一过程，可确定文本中知识实体的关联。结合待预测数据的实体标注预测与实体关系预测，即可识别出待预测数据的实体标注预测与实体关系预测中存在的实体及其相互关系，用于提取待预测数据的知识图谱三元组，从而实现自动化的知识抽取，且在保证精度的前提下有效提升了模型训练效率。

实施例二

请继续参阅图 2，本申请示出了一种基于 TextCNN 的知识抽取装置 10，以实施例一为基础，用以实现实施例一的基于 TextCNN 的知识抽取方法，其包括的各程序模块的功能：

10 在本实施例中，基于 TextCNN 的知识抽取装置 10 可以包括或被分割成一个或多个程序模块，一个或者多个程序模块被存储于存储介质中，并由一个或多个处理器所执行，以完成本申请，并可实现上述基于 TextCNN 的知识抽取方法。本申请所称的程序模块是指能够完成特定功能的一系列计算机程序指令段，比程序本身更适合于描述基于 TextCNN 的知识抽取装置在存储介质中的执行过程。以下描述将具体介绍本实施例各程序模块的功能：

15 本申请还提供了一种基于 TextCNN 的知识抽取装置 10，包括：

字向量字典构建模块 11：用于基于收集的第一训练数据，构建字向量字典；

词向量字典构建模块 12：用于基于收集的第一训练数据，构建词向量字典；

第一卷积神经网络构建及训练模块 13，用于构建第一卷积神经网络并基于第一优化算法训练第一卷积神经网络；第一卷积神经网络包括依次连接的第一嵌入层、第一多层卷积以及第一 softmax 函数；包括

20 字向量预处理单元，第二训练数据为预先标注数据，包括命名实体位置标签与命名实体关系标签，字向量预处理模块用于将第二训练数据分割为单一文字并去除特殊符号后，输入至第一嵌入层中；

25 字向量矩阵化单元，用于基于字向量字典，于第一嵌入层将字符级的第二训练数据进行字向量匹配，以将第二训练数据转化为矩阵形式；

第一多层卷积单元，用于对第一嵌入层输出的矩阵进行卷积运算，第一多层卷积单元包括位于前部的一组第一类一维卷积层以及位于后部的至少一组第二类一维卷积层，第一类一维卷积层包括长度数量不同、通道数量相同设置一维卷积核，第二类一维卷积层包括长度数量与通道数量均相同设置的一维卷积核，每一卷积层的数据均来自于前一卷积层输出，且卷积运算过程中，保持矩阵行数不变；

30 第一 softmax 函数输出单元，用于将第一多层卷积经由第一 softmax 函数进行输出，以确定每一文字的多个细分类别 BEMO 标注的预测概率；

第一卷积神经网络训练单元，用于根据 BEMO 标注预测概率与第二训练数据真实 BEMO 标签，计算损失函数交叉熵，并通过第一优化算法对损失函数进行最小化，以训练第一卷积神经网络；

35 第二卷积神经网络构建及训练模块 14，用于构建第二卷积神经网络并基于第二优化算法训练第二卷积神经网络，第二卷积神经网络包括依次连接的第二嵌入层、第二多层卷积单元、一池化层、两层全连接层以及第二 softmax 函数，包括：

词向量预处理单元：用于对第二训练数据进行分词，包括：

40 初步分词子单元，用于采用 jieba 库对第二训练数据进行初步分词，并比照第一卷积神经网络的预测进行修正，若初步分词结果与第一卷积神经网络的预测的分词结果不同，以第一卷积神经网络的预测的分词结果为准；

分词预处理子单元，用于去除初步分词中的特殊符号及非中文字符后，将处理后的第二训练数据输入至第二嵌入层；

45 字向量矩阵化单元，用于基于词向量词典，于第二嵌入层对分词后的第二训练数据进行词向量匹配，以将第二训练数据转化为矩阵形式；

- 第二多层卷积单元，用于基于第二多层卷积单元对第二嵌入层输出的矩阵进行卷积运算，第二多层卷积单元包括位于前部的一组第一类一维卷积层以及位于后部的至少一组第二类一维卷积层，第一类一维卷积层包括长度数量不同、通道数量相同设置一维卷积核，第二类一维卷积层包括长度数量与通道数量均相同设置的一维卷积核，每一卷积层的数据均来自于前一卷积层输出，且卷积运算过程中保持矩阵行数不变；
- 池化层，用于将第二多层卷积单元的输出输入至池化层进行压缩；
- 全连接层，用于将池化层的输出输入至两层全连接层中进行各个通道的信息融合；
- 第二 softmax 函数输出单元，用于将全连接层的输出输入至第二 softmax 函数中，确定多个实体关系标签的对应预测概率，
- 训练第二卷积神经网络单元，用于依据第二卷积神经网络输出的关系标签的预测概率与第二训练数据真实关系标签，计算第二损失函数交叉熵，并通过优化算法对损失函数进行最小化，以训练第二卷积神经网络；
- 知识图谱三元组提取模块 15，用于将待预测数据输入至训练完毕的第一卷积神经网络与第二卷积神经网络中，依据训练后第一卷积神经网络输出的实体标注预测与训练后第二卷积神经网络输出实体关系预测，提取待预测数据的知识图谱三元组：选取 BEMO 标注预测概率中最高概率值对应的类作为第一卷积神经网络输出的实体标注预测，选取预测概率值大于 0.5 的类作为第二卷积神经网络输出实体关系预测，以提取待预测数据的知识图谱三元组。
- 优选的，字向量字典构建模块 11 中，将收集的第一训练数据划分为单独文字并去除特殊符号与非中文字符后，代入 Word2Vec 算法进行训练，获取字向量并构建字向量字典；
- 优选的，词向量字典构建模块 12 中，将收集的第一训练数据进行分词并去除特殊符号与非中文字符后，代入 Word2Vec 算法进行训练，获取词向量字典。
- 进一步的，Word2Vec 算法训练通过 Python 中的 gensim 库实现。
- 优选的，基于 Python 中的 tensorflow 库建立第一卷积神经网络与第二卷积神经网络。
- 优选的，第一多层卷积包括 5 层卷积层，第一类一维卷积层包括了 3 类长度的一维卷积核，各对应 128 通道；四组的第二类一维卷积层的一维卷积核长度均为 3 且通道数为 384；和/或，第二多层卷积包括 3 层卷积层，第一类一维卷积层包括了 3 类长度的一维卷积核，各对应 128 通道；两组的第二类一维卷积层的一维卷积核长度均为 3 且通道数为 384。
- 优选的，第一优化算法与第二优化算法为 ADAM 算法或为 rmsprop 算法。
- 优选的，知识图谱三元组提取模块 15 中，若实体标注预测与实体关系预测结果相互矛盾，则放弃待预测数据的知识图谱三元组提取。
- 本申请所示的本申请提供的一种基于 TextCNN 知识抽取装置，通过卷积神经网络实现知识图谱中的知识抽取环节，在保证精度的前提下有效提升了模型训练效率。通过将训练文本转化为向量形式，并接入两类卷积神经网络模型（卷积层形式均为一维卷积核）中以提炼训练文本信息，分别实现命名实体识别以及实体关系识别。由于卷积神经网络具有并行计算特性，因此能够充分利用计算资源实现计算效率提升，训练完成的两类卷积神经网络模型通过预测融合实现自动化知识抽取。具体而言，通过将待预测数据分别转化为字向量与词向量形式，并接入至第一卷积神经网络与第二卷积神经网络处理，其中第一卷积神经网络用以实现命名实体识别。所用卷积神经网络为全卷积形式，其输入为字向量，输出为实体类别边界预测。经过这一过程可将原本连续的文本进行分词，并保留命名实体相关文字并进行实体归类；第二卷积神经网络实现知识抽取。所用卷积神经网络包括卷积层、池化层等，其输入包括字向量与词向量，输出为关系抽取识别。经过这一过程，可确定文本中知识实体的关联。结合待预测数据的实体标注预测与实体关系预测，即可识别出待预测数据的实体标注预测与实体关系预测中存在的实体及其相互关系，用于提取待预测数据的知识图谱三元组，从而实现自动化的知识抽取，且在保证精度的前提下有效提升了模型训练效率。

实施例三

本申请还提供一种计算机设备，如可以执行程序的智能手机、平板电脑、笔记本电脑、台式计算机、机架式服务器、刀片式服务器、塔式服务器或机柜式服务器（包括独立的服务器，或者多个服务器所组成的服务器集群）等。本实施例的计算机设备 20 至少包括但不限于：可通过系统总线相互通信连接的存储器 21、处理器 22，如图 3 所示。需要指出的是，图 3 仅示出了具有组件 21-22 的计算机设备 20，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多或者更少的组件。

本实施例中，存储器 21（即可读存储介质）包括闪存、硬盘、多媒体卡、卡型存储器（例如，SD 或 DX 存储器等）、随机访问存储器（RAM）、静态随机访问存储器（SRAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、可编程只读存储器（PROM）、磁性存储器、磁盘、光盘等。在一些实施例中，存储器 21 可以是计算机设备 20 的内部存储单元，例如该计算机设备 20 的硬盘或内存。在另一些实施例中，存储器 21 也可以是计算机设备 20 的外部存储设备，例如该计算机设备 20 上配备的插接式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡，闪存卡（Flash Card）等。当然，存储器 21 还可以既包括计算机设备 20 的内部存储单元也包括其外部存储设备。本实施例中，存储器 21 通常用于存储安装于计算机设备 20 的操作系统和各类应用软件，例如实施例一的基于 TEXTCNN 知识抽取装置 10 的程序代码等。此外，存储器 21 还可以用于暂时地存储已经输出或者将要输出的各类数据。

处理器 22 在一些实施例中可以是中央处理器（Central Processing Unit, CPU）、控制器、微控制器、微处理器、或其他数据处理芯片。该处理器 22 通常用于控制计算机设备 20 的总体操作。本实施例中，处理器 22 用于运行存储器 21 中存储的程序代码或者处理数据，例如运行基于 TEXTCNN 知识抽取装置 10，以实现实施例一的基于 TEXTCNN 知识抽取方法。

实施例四

本申请还提供一种计算机可读存储介质，如闪存、硬盘、多媒体卡、卡型存储器（例如，SD 或 DX 存储器等）、随机访问存储器（RAM）、静态随机访问存储器（SRAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、可编程只读存储器（PROM）、磁性存储器、磁盘、光盘、服务器、App 应用商城等等，其上存储有计算机程序，程序被处理器执行时实现相应功能。本实施例的计算机可读存储介质用于存储基于 TEXTCNN 知识抽取装置 10，被处理器执行时实现实施例一的基于 TEXTCNN 知识抽取方法。

上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。

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上述实施例方法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，但很多情况下前者是更佳的实施方式。

以上仅为本申请的优选实施例，并非因此限制本申请的专利范围，凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本申请的专利保护范围内