

权利要求书

1. 一种基于 TextCNN 的知识抽取方法, 其特征在于, 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;
- 5 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;
- 10 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
- 15 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 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
- 20 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;
- 25 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;
- 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
- 30 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;
- 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
- 35 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;
- 40 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;
- 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;
- 45 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

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。

- 5 2. 根据权利要求 1 所述的基于 TextCNN 的知识抽取方法,其特征在于,所述步骤 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;

- 10 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。

 3.根据权利要求 2 所述的基于 TextCNN 的知识抽取方法,其特征在于, Word2Vec 算法训练通过 Python 中的 gensim 库实现。

 4.根据权利要求 1 所述的基于 TextCNN 的知识抽取方法,其特征在于,基于 Python 中的 tensorflow 库建立所述第一卷积神经网络与所述第二卷积神经网络。

- 15 5.根据权利要求 1 所述的基于 TextCNN 的知识抽取方法,其特征在于,所述第一多层卷积包括 5 层卷积层,所述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;四组所述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,所述第二多层卷积包括 3 层卷积层,所述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;两组所述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。

 6.根据权利要求 1 所述的基于 TextCNN 的知识抽取方法,其特征在于,所述第一优化算法与所述第二优化算法为 ADAM 算法或为 rmsprop 算法。

- 25 7.根据权利要求 1 所述的基于 TextCNN 的知识抽取方法,其特征在于,所述步骤 S40 中,若所述实体标注预测与所述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。

 8.一种基于 TextCNN 的知识抽取装置,其特征在于,其包括:

 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;

- 30 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

 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;

- 35 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;

- 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;

 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;

- 45 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;

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;

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;

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;

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;

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;

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。

9.根据权利要求 8 所述的基于 TextCNN 的知识抽取装置,其特征在于,字向量字典构建模块中,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;

10.根据权利要求 9 所述的基于 TextCNN 的知识抽取装置,其特征在于,词向量字典构建模块中,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。

11.根据权利要求 8 所述的基于 TextCNN 的知识抽取装置,其特征在于,所述第一多层卷积包括 5 层卷积层,所述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;四组所述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,所述第二多层卷积包括 3 层卷积层,所述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;两组所述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。

12.一种计算机设备,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,其特征在于,所述处理器执行所述计算机程序时实现基于 TextCNN 的知识抽取方法的以下步骤:

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；

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；

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 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；

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；

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；

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；

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。

13. 根据权利要求 12 所述的计算机设备, 其特征在于, 所述步骤 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。

14. 根据权利要求 13 所述的计算机设备, 其特征在于, Word2Vec 算法训练通过 Python 中的 gensim 库实现。

15. 根据权利要求 12 所述的计算机设备, 其特征在于, 基于 Python 中的 tensorflow 库建立所述第一卷积神经网络与所述第二卷积神经网络。

16. 根据权利要求 12 所述的计算机设备, 其特征在于, 所述第一多层卷积包括 5 层卷积层, 所述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; 四组所述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, 所述第二多层卷积包括 3 层卷积层, 所述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; 两组所述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。

17. 一种计算机可读存储介质, 其上存储有计算机程序, 其特征在于, 所述计算机程序被处理器执行时实现基于 TextCNN 的知识抽取方法的以下步骤:

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;

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;

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 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;

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;

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;

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;

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。

18. 根据权利要求 17 所述的计算机可读存储介质, 其特征在于, 所述步骤 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。

19. 根据权利要求 17 所述的计算机可读存储介质, 其特征在于, 基于 Python 中的 tensorflow 库建立所述第一卷积神经网络与所述第二卷积神经网络。

20. 根据权利要求 17 所述的基于 TextCNN 的知识抽取方法, 其特征在于, 所述第一多层卷积包括 5 层卷积层, 所述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; 四组所述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, 所述第二多层卷积包括 3 层卷积层, 所述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; 两组所述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。