(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 112487823 A
(43)申请公布日 2021.03.12

(21)申请号 202011294655.6
(22)申请日 2020.11.18
(71)申请人 广东电力信息科技有限公司
地址 510030 广东省广州市越秀区东风东路808号509房
(72)发明人 郑颖龙 周昉昉 刘佳木 赖蔚蔚 吴广财 郑杰生 林嘉鑫 叶杭
(74)专利代理机构 北京世誉鑫诚专利代理有限公司 11368
代理人 任欣生
(51)Int.Cl.
G06F 40/30(2020.01)
G06F 40/284(2020.01)
G06F 40/289(2020.01)
G06K 9/62(2006.01)
G06N 3/04(2006.01)

权利要求书1页 说明书3页

(54)发明名称

一种基于BERT模型的文本语义相似度计算方法

(57)摘要

本发明公开的基于BERT模型的文本语义相似度计算方法,通过对用户输入的两个句子做子词切分,得到两个子词序列,分别在两个子词序列的头部、连接处及尾部设置标记,得到完整的子词序列,将子词序列输入BERT模型,得到子词序列中各个子词对应的语义向量,将头部特殊标记对应的语义向量输入神经网络模型的全连接层,得到维度为2的语义向量,将维度为2的语义向量输入神经网络模型的Softmax层做归一化,得到两个句子相似的概率和不相似的概率,根据两个句子相似的概率和不相似的概率,确定两个句子的语义相似度,避免了因分词可能引入的错误,能够考虑文本的上下文语义,提高了语义相似度计算的精确度。

CN 112487823 A

1.一种基于BERT模型的文本语义相似度计算方法,其特征在于,包括:

对用户输入的两个句子做子词切分,得到两个子词序列;

分别在所述两个子词序列的头部、连接处及尾部设置标记,得到完整的子词序列;

将所述子词序列输入BERT模型,得到所述子词序列中各个子词对应的语义向量;

将所述头部标记对应的语义向量输入神经网络模型的全连接层,得到维度为2的语义向量,其中,所述维度为2的语义向量分别表示两个句子相似和不相似;

将所述维度为2的语义向量输入神经网络模型的Softmax层做归一化,得到两个句子相似的概率和不相似的概率;

根据所述两个句子相似的概率和不相似的概率,确定所述两个句子的语义相似度。

2.根据权利要求1所述的基于BERT模型的文本语义相似度计算方法,其特征在于,根据所述两个句子相似的概率和不相似的概率,确定所述两个句子的语义相似度包括:

判断相似的概率是否大于不相似的概率,若是,则确定两个句子相似并将相似的概率作为两个句子的语义相似度,若否,则确定两个句子不相似。

3.一种计算机程序产品,其特征在于,所述计算机程序产品包括存储在非暂态计算机可读存储介质上的计算机程序,所述计算机程序包括程序指令,当所述程序指令被计算机执行时,使所述计算机执行如权利要求1-2所述的方法。

4.一种非暂态计算机可读存储介质,其特征在于,所述非暂态计算机可读存储介质存储计算机指令,所述计算机指令使所述计算机执行如权利要求1-2所述的方法。

一种基于BERT模型的文本语义相似度计算方法

技术领域

[0001] 本发明涉及文本处理技术领域，具体涉及一种基于BERT模型的文本语义相似度计算方法。

背景技术

[0002] 语义相似度计算是人工智能自然语言处理领域的基础任务之一，是文本查重、智能问答等上层应用的基础支撑技术。语义相似度意在对于给定的两个文本，从语义的角度度量二者之间的相似性，通常会给出一个0到1之间的语义相似度分值，分值越高代表越相似。

[0003] 现有的语义相似度方案有的基于字面进行计算，无法考虑语义上的相似性。有的方案基于Word2Vec等静态词向量计算语义相似度，无法考虑一词多义的情况，另外由于需要先进行分词，可能存在分词错误的情况，导致语义相似度计算的精确度较低。

发明内容

[0004] 为解决现有技术的不足，本发明实施例提供了一种基于BERT模型的文本语义相似度计算方法，该方法包括以下步骤：

[0005] 对用户输入的两个句子做子词切分，得到两个子词序列；

[0006] 分别在所述两个子词序列的头部、连接处及尾部设置标记，得到完整的子词序列；

[0007] 将所述子词序列输入BERT模型，得到所述子词序列中各个子词对应的语义向量；

[0008] 将所述头部特殊标记对应的语义向量输入神经网络模型的全连接层，得到维度为2的语义向量，其中，所述维度为2的语义向量分别表示两个句子相似和不相似；

[0009] 将所述维度为2的语义向量输入神经网络模型的Softmax层做归一化，得到两个句子相似的概率和不相似的概率；

[0010] 根据所述两个句子相似的概率和不相似的概率，确定所述两个句子的语义相似度。

[0011] 优选地，根据所述两个句子相似的概率和不相似的概率，确定所述两个句子的语义相似度包括：

[0012] 判断相似的概率是否大于不相似的概率，若是，则确定两个句子相似并将相似的概率作为两个句子的语义相似度，若否，则确定两个句子不相似。

[0013] 本发明实施例提供的基于BERT模型的文本语义相似度计算方法，具有以下有益效果：

[0014] 将BERT模型应用于计算文本语义相似度，能够达到更好的语义建模效果，基于字符计算语义相似度，不依赖分词，避免了因分词可能引入的错误，能够考虑文本的上下文语义，提高了语义相似度计算的精确度。

具体实施方式

[0015] 以下结合具体实施例对本发明作具体的介绍。

[0016] 本发明提供的实施例提供的基于BERT模型的文本语义相似度计算方法，包括以下步骤：

[0017] S101，对用户输入的两个句子做子词切分，得到两个子词序列。

[0018] 其中，每个汉字都是一个子词，一个英文单词可能会被切分成多个子词。

[0019] S102，分别在所述两个子词序列的头部、连接处及尾部设置标记，得到完整的子词序列。

[0020] 作为本发明一个具体的实施例，对于A1、A2...An和B1、B2...Bm两个子词序列，得到的完整的子词序列为[CLS]、A1、A2...An、[sep]、B1、B2...Bm、[sep]。

[0021] S103，将所述子词序列输入BERT模型，得到所述子词序列中各个子词对应的语义向量。

[0022] S104，将头部特殊标记对应的语义向量输入神经网络模型的全连接层，得到维度为2的语义向量，其中，维度为2的语义向量分别表示两个句子相似和不相似。

[0023] S105，将维度为2的语义向量输入神经网络模型的Softmax层做归一化，得到两个句子相似的概率和不相似的概率。

[0024] S106，根据两个句子相似的概率和不相似的概率，确定两个句子的语义相似度。

[0025] 可选地，根据所述两个句子相似的概率和不相似的概率，确定所述两个句子的语义相似度包括：

[0026] 判断相似的概率是否大于不相似的概率，若是，则确定两个句子相似并将相似的概率作为两个句子的语义相似度，若否，则确定两个句子不相似。

[0027] 本发明实施例提供的基于BERT模型的文本语义相似度计算方法，通过对用户输入的两个句子做子词切分，得到两个子词序列，分别在两个子词序列的头部、连接处及尾部设置标记，得到完整的子词序列，将子词序列输入BERT模型，得到子词序列中各个子词对应的语义向量，将头部特殊标记对应的语义向量输入神经网络模型的全连接层，得到维度为2的语义向量，将维度为2的语义向量输入神经网络模型的Softmax层做归一化，得到两个句子相似的概率和不相似的概率，根据两个句子相似的概率和不相似的概率，确定两个句子的语义相似度，避免了因分词可能引入的错误，能够考虑文本的上下文语义，提高了语义相似度计算的精确度。

[0028] 在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部分，可以参见其他实施例的相关描述。

[0029] 可以理解的是，上述方法及装置中的相关特征可以相互参考。

[0030] 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统，装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0031] 在此提供的算法和显示不与任何特定计算机、虚拟系统或者其它设备固有相关。各种通用系统也可以与基于在此的示教一起使用。根据上面的描述，构造这类系统所要求的结构是显而易见的。此外，本发明也不针对任何特定编程语言。应当明白，可以利用各种编程语言实现在此描述的本发明的内容，并且上面对特定语言所做的描述是为了披露本发明的最佳实施方式。

[0032] 此外，存储器可能包括计算机可读介质中的非永久性存储器，随机存取存储器(RAM)和/或非易失性内存等形式，如只读存储器(ROM)或闪存(flash RAM)，存储器包括至少一个存储芯片。

[0033] 本领域内的技术人员应明白，本申请的实施例可提供为方法、系统、或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请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、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0034] 在一个典型的配置中，计算设备包括一个或多个处理器(CPU)、输入/输出接口、网络接口和内存。

[0035] 存储器可能包括计算机可读介质中的非永久性存储器，随机存取存储器(RAM)和/或非易失性内存等形式，如只读存储器(ROM)或闪存(flash RAM)。存储器是计算机可读介质的示例。

[0036] 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带，磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括暂存电脑可读媒体(transitory media)，如调制的数据信号和载波。

[0037] 还需要说明的是，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，并不排除在包括要素的过程、方法、商品或者设备中还存在另外的相同要素。

[0038] 本领域技术人员应明白，本申请的实施例可提供为方法、系统或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且，本申请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、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0039] 以上仅为本申请的实施例而已，并不用于限制本申请。对于本领域技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原理之内所作的任何修改、等同替换、改进等，均应包含在本申请的权利要求范围之内。