(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111581351 B
(45)授权公告日 2023.05.02

(21)申请号 202010367701.4

(22)申请日 2020.04.30

(65)同一申请的已公布的文献号
申请公布号 CN 111581351 A

(43)申请公布日 2020.08.25

(73)专利权人 识因智能科技(北京)有限公司
地址 102600 北京市大兴区宏福路8号1层133室

(72)发明人 王春辉 胡勇

(74)专利代理机构 北京中北知识产权代理有限公司 11253
专利代理师 卢业强

(51)Int.Cl.
G06F 16/33(2019.01)
G06F 40/30(2020.01)
G06N 3/0464(2023.01)
G06N 3/08(2023.01)
G06N 3/048(2023.01)

(56)对比文件
CN 108399163 A,2018.08.14
CN 109635109 A,2019.04.16
CN 109948165 A,2019.06.28
US 2018349359 A1,2018.12.06
WeijiangLi.Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification.《Neurocomputing》.2020,全文.
Zhuosheng Zhang.Effective Subword Segmentation for Text Comprehension.《IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 27, Issue: 11, November 2019)》.2019,全文.
凡子威;张民;李正华.基于BiLSTM并结合自注意力机制和句法信息的隐式篇章关系分类.计算机科学.2019,(05),全文.

审查员 邓成

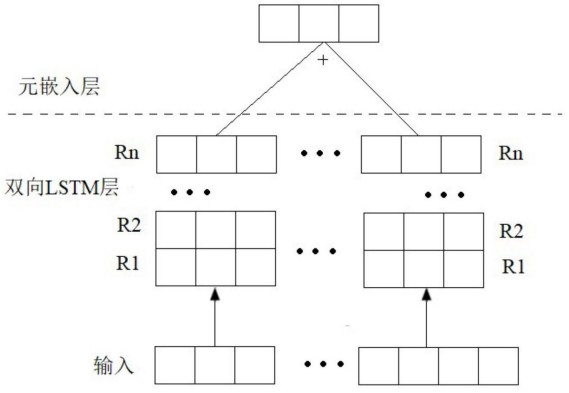
权利要求书1页 说明书4页 附图1页

(54)发明名称

一种基于多头自注意力机制的动态元嵌入方法

(57)摘要

本发明公开一种基于多头自注意力机制的动态元嵌入方法。所述方法包括:将输入句子中的每个词表示为词向量序列,将每个词向量映射到同一维度,基于多头自注意力机制计算嵌入矩阵,得到元嵌入表示的词向量矩阵。本发明利用多头自注意力机制进行多次计算,解决了现有DME、CDME动态元嵌入方法使用句子级别的自注意力确定不同嵌入集的权重,其中自注意力方法着重于学习各个词向量线性回归的参数,缺乏多角度的综合考虑,容易造成元嵌入权重的学习不充分的问题。



CN 111581351 B

1.一种基于多头自注意力机制的动态元嵌入方法,其特征在于,包括以下步骤:

步骤1,将输入句子中的每个词表示为词向量序列$\{w_{i,j}\}_{j=1}^{s}$,$w_{i,j}$为嵌入第i个嵌入集的第j个词,j=1,2,…,s,s为句子中词的数量,i=1,2,…,n,n为嵌入集的数量;

步骤2,通过一个全连接层将每个词向量映射到同一维度,表示为:

$$w'_{i,j}=p_i w_{i,j}+c_i \quad (1)$$

其中,p_i、c_i为学习参数;

步骤3,基于多头自注意力机制计算嵌入矩阵,按(2)～(5)式得到元嵌入表示的词向量矩阵$B=[w''_{i,j}]_{n\times s}$:

$$w''_{i,j}=A_{1\times R}X_{R\times 1}w'_{i,j}=\sum_{r=1}^{R}a_{i,j}^{r}x_r w'_{i,j} \quad (2)$$

$$A_{1\times R}=(a_{i,j}^{1},a_{i,j}^{2},\cdots,a_{i,j}^{R}) \quad (3)$$

$$a_{i,j}^{r}=\phi\left(a_r w'_{i,j}+b_r\right) \quad (4)$$

$$X_{R\times 1}=(x_1,x_2,\cdots,x_R)^{T} \quad (5)$$

其中,a_r、b_r和x_r为学习参数,r=1,2,…,R,R为多头自注意力机制的计算次数,$A_{1\times R}$为嵌入矩阵,φ为softmax或sigmoid门控函数。

2.根据权利要求1所述的基于多头自注意力机制的动态元嵌入方法,其特征在于,所述方法还包括降低输入句子噪声的预处理步骤。

3.根据权利要求1所述的基于多头自注意力机制的动态元嵌入方法,其特征在于,所述方法还包括将训练模型的损失函数增加一个惩罚项NewPT:

$$NewPT=\left\|A_{1\times R}^{T}A_{1\times R}-I\right\|_F^2 \quad (6)$$

其中,$\|\bullet\|_F^2$表示求矩阵的Frobenius范数,I为单位矩阵。

一种基于多头自注意力机制的动态元嵌入方法

技术领域

[0001] 本发明属于自然语言理解领域,具体涉及一种基于多头自注意力机制的动态元嵌入方法。

背景技术

[0002] 深度学习中的词向量(单词的分布式表示,也称为词嵌入)对自然语言处理的许多任务中都有应用。近年来,Word2Vec、GloVe等预训练嵌入集得到了广泛应用。元嵌入学习是是集成词嵌入的一种技术,目的是将同一个词的不同词嵌入通过某种方式融合得到新的词向量表示。元嵌入学习得到的元嵌入捕获了不同嵌入集中词法语义的互补信息。

[0003] 元嵌入学习包括静态元嵌入和动态元嵌入。静态元嵌入把元嵌入学习作为预处理过程。CONC、SVD、1TON和1TON+是常用的四种基线静态元嵌入学习方法。前三种方法在嵌入集的重叠词汇上学习元嵌入。CONC串联来自不同嵌入集的单词向量。SVD在CONC的基础上执行降维操作。1TON假设存在该单词的元嵌入,比如一开始随机初始化元嵌入,并使用该元嵌入通过线性投影预测单个词向量集中该单词的表示,进行了微调的元嵌入期望包含来自所有嵌入集的知识。在静态元嵌入学习中,会遇到这样的未登录词问题:单词A在嵌入集M中出现,但是在嵌入集N中没有录入。为了解决未登录词问题,1TON+首先随机初始化OOV(Out-of-vocabulary)和元嵌入的向量表示,然后使用类似于1TON的预测设置来更新元嵌入和OOV嵌入。因此,1TON+同时达到两个目标:学习元嵌入和扩展词汇表(最终会是所有嵌入集词汇表的并集)。动态元嵌入将集成词向量的过程融入到特定NLP(Natural Language Processing,自然语言处理)任务端到端模型的过程中,使得模型可以根据特定任务自主选择不同词向量的权重。将元嵌入思想应用于句子表示,可以动态地学习不同嵌入集的注意力权重。计算权重的基本框架包括元嵌入层、句子编码层、匹配层和分类器。嵌入层采用DME(dynamic meta-embeddings)算法,利用自注意力机制和门控函数,动态计算集成各个嵌入集的权重;也可采用上下文相关的CDME(contextualized DME)算法来增强投影嵌入,用双向的长短时记忆网络LSTM(Long Short-Term Memory)替代简单的线性映射。

[0004] DME和CDME动态元嵌入算法,使用句子级别的自注意力来确定不同嵌入集的权重,其中自注意力方法着重于学习各个词向量线性回归的参数,缺乏多角度的综合考虑,很容易造成元嵌入权重的学习不充分。

发明内容

[0005] 为了解决现有技术中存在的上述问题,本发明提出一种基于多头自注意力机制的动态元嵌入方法。

[0006] 为实现上述目的,本发明采用如下技术方案:

[0007] 一种基于多头自注意力机制的动态元嵌入方法,包括以下步骤:

[0008] 步骤1,将输入句子中的每个词表示为词向量序列$\{w_{i,j}\}_{j=1}^{s}$,$w_{i,j}$为嵌入第i个嵌入集的第j个词,j=1,2,…,s,s为句子中词的数量,i=1,2,…,n,n为嵌入集的数量;

[0009] 步骤2,通过一个全连接层将每个词向量映射到同一维度,表示为:

[0010] $w'_{i,j}=p_i w_{i,j}+c_i$ (1)

[0011] 其中,p_i、c_i为学习参数;

[0012] 步骤3,基于多头自注意力机制计算嵌入矩阵,按(2)～(5)式得到元嵌入表示的词向量矩阵$B=[w''_{i,j}]_{n\times s}$:

[0013] $$w''_{i,j}=A_{1\times R}X_{R\times 1}w'_{i,j}=\sum_{r=1}^{R}a^r_{i,j}x_r w'_{i,j} \quad (2)$$

[0014] $$A_{1\times R}=(a^1_{i,j},a^2_{i,j},\cdots,a^R_{i,j}) \quad (3)$$

[0015] $$a^r_{i,j}=\phi\left(a_r w'_{i,j}+b_r\right) \quad (4)$$

[0016] $X_{R\times 1}=(x_1,x_2,\cdots,x_R)^T$ (5)

[0017] 其中,a_r、b_r和x_r为学习参数,r=1,2,…,R,R为多头自注意力机制的计算次数,$A_{1\times R}$为嵌入矩阵,Φ为softmax或sigmoid门控函数。

[0018] 与现有技术相比,本发明具有以下有益效果:

[0019] 本发明通过将输入句子中的每个词表示为词向量序列,将每个词向量映射到同一维度,基于多头自注意力机制计算嵌入矩阵,得到元嵌入表示的词向量矩阵,实现了词向量序列的动态嵌入。本发明由于利用多头自注意力机制进行多次计算,解决了现有动态元嵌入(DME、CDME)使用句子级别的自注意力确定不同嵌入集的权重,其中自注意力方法侧重学习各个词向量线性回归的参数,缺乏多角度的综合考虑,容易造成元嵌入权重的学习不充分的问题。

附图说明

[0020] 图1为基于多头自注意力机制进行动态元嵌入的结构框图。

具体实施方式

[0021] 下面结合附图1对本发明作进一步详细说明。

[0022] 本发明实施例一种基于多头自注意力机制的动态元嵌入方法,包括以下步骤:

[0023] S101、将输入句子中的每个词表示为词向量序列$\{w_{i,j}\}_{j=1}^{s}$,$w_{i,j}$为嵌入第i个嵌入集的第j个词,j=1,2,…,s,s为句子中词的数量,i=1,2,…,n,n为嵌入集的数量;

[0024] S102、通过一个全连接层将每个词向量映射到同一维度,表示为:

[0025] $w'_{i,j}=p_i w_{i,j}+c_i$ (1)

[0026] 其中,p_i、c_i为学习参数;

[0027] S103、基于多头自注意力机制计算嵌入矩阵,按(2)～(5)式得到元嵌入表示的词向量矩阵$B=[w''_{i,j}]_{n\times s}$:

[0028] $$w''_{i,j}=A_{1\times R}X_{R\times 1}w'_{i,j}=\sum_{r=1}^{R}a^r_{i,j}x_r w'_{i,j} \quad (2)$$

[0029] $$A_{1\times R}=(a^1_{i,j},a^2_{i,j},\cdots,a^R_{i,j}) \quad (3)$$

[0030] $$a_{i,j}^{r}=\phi\left(a_{r}w_{i,j}^{'}+b_{r}\right) \quad (4)$$

[0031] $X_{R\times1}=(x_1,x_2,\cdots,x_R)^T$ (5)

[0032] 其中，a_r、b_r和x_r为学习参数，r＝1,2,…,R,R为多头自注意力机制的计算次数(即头数)，$A_{1\times R}$为嵌入矩阵(也称多头自注意力矩阵)，φ为softmax或sigmoid门控函数。

[0033] 在本实施例中，步骤S101主要用于将输入句子中的每个词表示为词向量序列。词向量序列指的是同一个词在n个嵌入集上的词嵌入集合，例如，将单词“USE”嵌入n＝3个嵌入集GloVe、fastText和word2vec，用GloVe得到w_1，用fastText得到w_2，用word2vec得到w_3，则W_(USE)＝{w_1,w_2,w_3}，需要对这个词向量序列做元嵌入。

[0034] 步骤S102主要用于将每个词向量映射到同一维度。根据(1)式，如果当前词向量$w_{i,j}$的维度为d×1，p_i的维度为d′×d，c_i是的维度为d′×1，则映射后$w'_{i,j}$的维度为d′×1。维度值一般取200～300比较适宜。

[0035] 步骤S103主要用于基于多头自注意力机制实现动态元嵌入。多头自注意力机制是自注意力机制的拓展，多头就是指计算多次。每个自注意力计算，首先经过简单的线性映射和tanh激活函数，再用softmax或者sigmoid门控函数进行降维，得到一个求和权重向量。经过多次的自注意力计算后得到嵌入矩阵，从而实现基于多头自注意力机制的动态元嵌入，得到元嵌入表示向量。得到元嵌入表示向量后，将元嵌入作为每个词的表示输入到上层神经网络(如编码层、匹配层和分类器)，最终通过分类器映射到维度为标签数的one-hot向量。

[0036] 本实施例采用多头自注意力机制，通过多次不同角度的自注意力计算，可以学习到互补的权重参数，使元嵌入权重的学习会更加充分，因此，可以解决现有的DME、CDME动态元嵌入方法使用句子级别的自注意力确定不同嵌入集的权重，由于其中自注意力方法侧重学习各个词向量线性回归的参数，缺乏多角度的综合考虑，容易造成元嵌入权重的学习不充分的问题。

[0037] 作为一种可选实施例，所述方法还包括降低输入文本噪声的预处理步骤。

[0038] 在本实施例中，通过预处理降低输入文本的噪声。降噪方法一般包括：去除干扰的标点符号，去除停用词，字母强制小写，限定句子最大长度。

[0039] 作为一种可选实施例，所述方法还包括将训练模型的损失函数增加一个惩罚项NewPT：

[0040] $$NewPT=\left\|A_{1\times R}^{T}A_{1\times R}-I\right\|_{F}^{2} \quad (6)$$

[0041] 其中，$\|\bullet\|_F^2$表示求矩阵的Frobenius范数，I为单位矩阵。

[0042] 在本实施例中，由于多头自注意力机制始终提供R个相似的求和权重($a_1,a_2,\cdots,a_R$)，则嵌入矩阵可能会出现冗余问题。因此，训练模型的目标函数需要在原损失项的基础上迭加一个惩罚项来鼓励不同关注点之间求和权重向量的多样性。评估多样性的一般方法是利用任意两个求和权重向量之间的Kullback Leibler差异(以下简称KL散度)，KL＝$a_i\log_e(a_i/a_j)$。根据KL的表达式可以看出，此方法没有对($a_1,a_2,\cdots,a_R$)的差异性进行约束，以包含KL的目标函数学习得到的参数，很可能形成冗余，也就是说($a_1,a_2,\cdots,a_R$)很可能都是相似的。因此，KL惩罚项不能使每个权重向量都能专注语义的一个方面。为此，本实施例引

入一个区别于KL的新的惩罚项，使用嵌入矩阵的转置与嵌入矩阵的积$A_{l\times R}^{T}A_{l\times R}$减去单位矩阵I后的Frobenius范数，作为冗余的度量，见(6)式。这样可以使不同的权重向量关注的部分不一样。$A_{l\times R}^{T}A_{l\times R}$减去单位矩阵I后，对角线上的特征元素近似为1，这样可以鼓励每个求和向量集中在尽可能少的特征元素上，从而迫使每个向量集中于一个方面，而所有其它元素都设为0，这将惩罚不同求和向量之间的冗余。

[0043] 下面给出本发明所述方法及现有的DME、CDME动态嵌入方法，应用于识别两个句子是否表达相同的含义时的一组实验数据。表1是在NLI数据集(SNLI Dev，SNLI Test，MultiNLI mismatched，MultiNLI matched)及全集AllNLI上的实验结果，表中的数据是识别的准确率。

[0044] 表1 NLI数据集上的实验结果对比

[0045]

	SNLI Dev	SNLI Test	MultiNLI mismatched	MultiNLI matched	AllNLI
DME	86.7%	86.2%	74.3%	73.5%	80.3%
CDME	87.1%	86.4%	74.3%	74.2%	80.5%
本发明	87.4%	87.1%	74.7%	74.7%	81.0%

[0046] 由表1的结果可以看出，本发明在自然语言推理任务中，在NLI数据集上的识别准确率优于现有DME、CDME动态嵌入方法。在全集AllNLI上也有很好的表现。

[0047] 上述仅对本发明中的几种具体实施例加以说明，但并不能作为本发明的保护范围，凡是依据本发明中的设计精神所做出的等效变化或修饰或等比例放大或缩小等，均应认为落入本发明的保护范围。

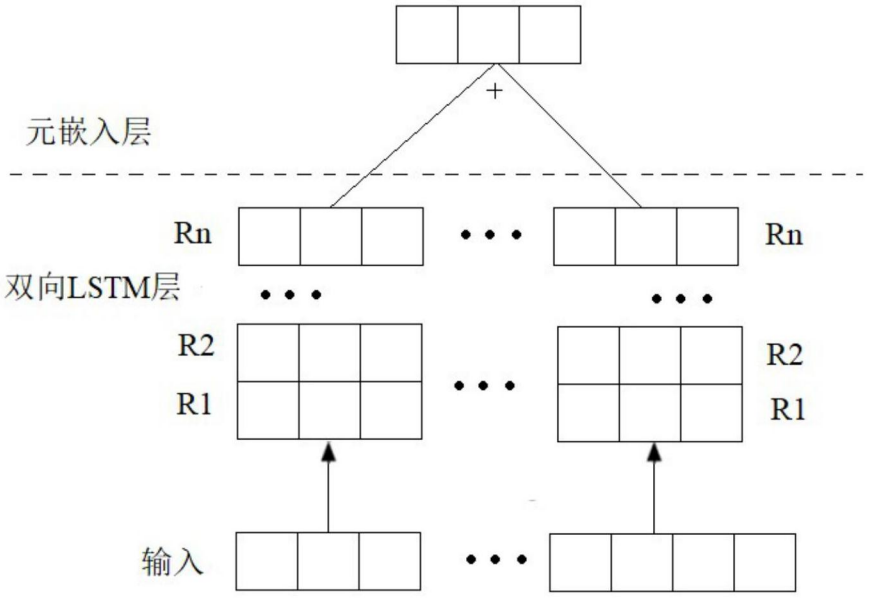


图1