(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110851593 B
(45)授权公告日 2024.01.05

(21)申请号 201910898057.0

(22)申请日 2019.09.23

(65)同一申请的已公布的文献号
申请公布号 CN 110851593 A

(43)申请公布日 2020.02.28

(73)专利权人 天津大学
地址 300072 天津市南开区卫津路92号

(72)发明人 赵东浩 张鹏

(74)专利代理机构 天津市北洋有限责任专利代理事务所 12201
专利代理师 韩帅

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 18/214(2023.01)
G06F 18/21(2023.01)
G06N 3/045(2023.01)
G06N 3/0464(2023.01)
G06N 3/0442(2023.01)
G06N 3/084(2023.01)
G06F 16/383(2019.01)
G06F 16/387(2019.01)
G06F 40/289(2020.01)

(56)对比文件
CN 106776581 A,2017.05.31
CN 108363714 A,2018.08.03
CN 108363769 A,2018.08.03
CN 109522548 A,2019.03.26
US 2003216919 A1,2003.11.20
胡朝举;赵晓伟.基于词向量技术和混合神经网络的情感分析.计算机应用研究.2017,(第12期),全文.
郭文姣;欧阳昭连;李阳;郭柯磊;杜然然;池慧.应用共词分析法揭示生物医学工程领域的研究主题.中国生物医学工程学报.2012,(第04期),全文.

审查员 支玉亮

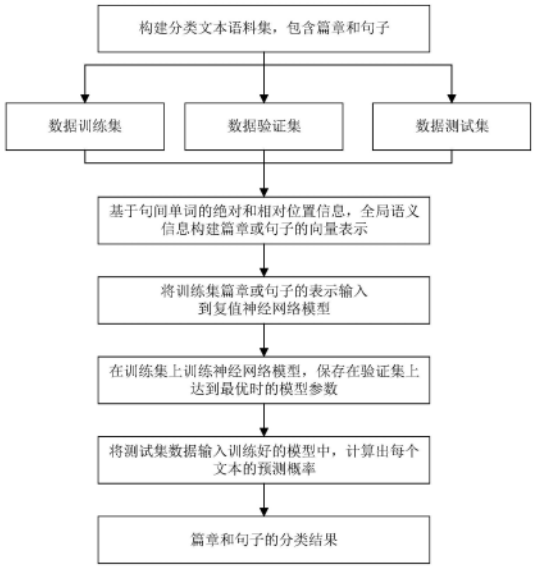
权利要求书2页 说明书5页 附图3页

(54)发明名称

一种基于位置与语义的复值词向量构建方法

(57)摘要

本发明公开了一种基于位置与语义的复值词向量构建方法,包括以下步骤:搜集文本分类语料集,并将其分为训练集,验证集和测试集;对语料集中的文本进行预处理(去除停止词);运用相对位置信息和全局语义信息,构建句子表示;将训练语料集的词向量输入到复值神经网络中,训练出语义分类模型;将验证集文本词向量输入到复值神经网络模型中,从而计算出每个样本的预测概率;并将基于验证集得出的模型在测试集上测试;本发明克服了文本分类语料集相对缺乏的现状,能够更充分的提取文本的特征信息(位置信息),融合文本的位置信息与全局语义信息,并将复值词向量应用到复制神经网络,使这些神经网络模型有较强的判别能力。



CN 110851593 B

1.一种基于位置与语义的复值词向量构建方法,其特征在于,包括如下步骤:

(1)采用jieba分词工具,对篇章和句子进行分词,从而构建多模态分类语料集;

(2)从(1)构建的多模态分类语料集中,随机选取80%*N个样本作为训练集,10%*N个样本划分为验证集和剩余的10%*N个样本划分为测试集,并分别对训练集、验证集和测试集进行预处理;

(3)选取多模态分类语料集中预处理之后的句子构建复值神经网络模型,进而建立复值神经网络模型损失函数:

$$loss = \sum_{i=1}^{N}[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)]$$

其中:y_i代表真实类别标签,$\hat{y}_i$代表预测结果;其中:

3.1、根据单词在篇章和句子的位置构造每个词的绝对位置索引,即:

$p_i = e^{i\beta POS}$

$$\beta = \frac{2\pi}{p_{j,i}}$$

其中:p_i表示当前词的位置向量,POS为其绝对位置,β为其初始化的周期,i为复数的虚部表示;

3.2、运用glove工具得到每个篇章或句子中单词的300维词向量w_i,同时建立一个位置信息的矩阵,每一个位置索引对应一个300维的位置向量p_i进而构造每个词的相对位置索引,即为篇章和句子中的每个词的表示:

$x_i = w_i e^{i\beta POS}$

所述步骤3.2中每个词的相对位置索引是通过如下公式获得篇章或句子的向量表示:

单词维度间相对位置:

$$p_{(j,pos+n)} = w_{j,k} e^{\frac{2\pi(pos+n)}{p_{j,k}} i}$$

$$= w_{j,k} e^{\frac{2\pi(pos)}{p_{j,k}} i} \times e^{\frac{2\pi n}{p_{j,k}} i}$$

$$= x(j,pos) \times e^{\frac{2\pi n}{p_{j,k}} i}$$

其中:x表示单词的向量表示,$w_{j,k}$表示单词的语义向量,n表示单词间隔;

单词间相对位置:

$$p_{(j_1,pos+n)} = w_{j_1,k} e^{\frac{2\pi(pos+n)}{p_{\cdot,k}} i}$$

$$= w_{j_1,k} e^{\frac{2\pi(pos)}{p_{\cdot,k}} i} \times \frac{w_{j_2,k}}{w_{j_1,k}} e^{\frac{2\pi n}{p_{\cdot,k}} i}$$

$$= x(j_2,pos) \times \frac{w_{j_2,k}}{w_{j_1,k}} e^{\frac{2\pi n}{p_{\cdot,k}} i}$$

其中:p.k表示不同位置单词的第k个维度;

3.3、文本的每个词的词向量连同它的相对位置信息向量按其在句子中的顺序输入到

Fasttext和LSTM,CNN网络中,具体计算公式如下:

Fasttext:

$Z^C=\sigma(Ax-By+b_r)+i\sigma(Bx-Ay+b_i)$

CNN

$Z^C=Ax-By+i(Bx+Ay)$

其中,σ表示sigmod激活函数,A和B分别表示权重的实部和虚部,x和y表示输入特征的实部和虚部;

RNN

$$h_t^C = f(W^C h_{t-1}^C + V^C x_t^C + b^C)$$

其中,x_t表示每个词的输入,$\vec{h}_t$表示分别得到的网络隐层状态;RNN公式中权重与输入,输出均使用了复值表示;通过用上述的公式,输入每个词的表示300维x_t,分别得到网络隐层状态表示128维$\vec{h}_t$;

3.4、将上述步骤中最终的网络隐层输出$\vec{h}_t$输入到一个非线性全连接神经网络层得到神经网络表示向量x,再将表示x输入到softmax分类层输出最终的类别y:

$$\boldsymbol{x} = tanh(\mathbf{W}_R \vec{\boldsymbol{h}}_t + \boldsymbol{b}_R)$$

$y=\mathrm{softmax}(W_s x+b_s)$;

(4)将复值神经网络模型在训练集上进行训练获得语义分类模型;

(5)在验证集对训练后神经网络模型进行效果验证,并记录保存达到最优时的模型参数;

(6)用上一步中保存的最优的模型参数去测试测试集上的样本,最终得到每个测试样本的预测结果,对比测试标签,计算出分类准确率。

一种基于位置与语义的复值词向量构建方法

技术领域

[0001] 本发明涉及文本分类技术领域，特别涉及一种基于位置与语义的复值词向量构建方法。

背景技术

[0002] 在过去几年随着科学技术的迅猛发展，特别是互联网和社交网络的快速发展，各种信息充斥在互联网上，这其中就包括用户在社交平台上发表的评论和自己的某些观点，这也成为用户日常生活中获取信息的主要来源之一。人们可能通过互联网获取大量资料，但如何对这些大量资料做出合理有效的管理，越来越成为人们所关心的问题。对大量信息一种很常见的管理方式就是分类，由此可见文本分类蕴含巨大的社会价值。本发明主要研究句子和篇章的情感或所属类别。

[0003] 分类任务在自然语言处理任务中扮演重要角色。分类任务可简单分为二分类(垃圾邮件分类等等)，多分类(文本的情感状态)，它的发展广泛受到业界和学术界的关注。本发明不仅讨论了句子的情感，还对句子所属类别进行了判断，这是文本分类领域的细粒度任务。例如“是不是看起来不可能:这让连环杀手杰弗里·达默很无聊”。本句的情感是消极情感，主要由单词“无聊”所决定。

[0004] 基于位置与语义的复值向量表示的神经网络分类方法旨在区分给定句子的情感极性或所属分类。当然目前工业界和学术界都意识到句子中单词情感信息的重要性，并试图通过设计一系列分类模型来更好的区分它们。然而当前的方法通常忽略了句中单词位置信息的重要性，我们并不知道单词所在位置是语义信息重要还是位置信息重要，当单词顺序发生改变但词没有变时，我们期望模型能够更好的识别出单词语义发生了改变。因此本发明关注到句中单词顺序与语义的关系吗，构建了复值词向量encode词间位置和语义信息，并通过复值神经网络模型提取分类信息。

[0005] 现在，基于复值神经网络模型已经被研究者们用来建模一些自然语言处理任务，并且已经非常成功。然而，目前的方法仅仅使用了复数向量，没有更好的利用复值词向量，挖掘句中单词的相对位置信息。

发明内容

[0006] 本发明所要解决的技术问题是克服现有技术的不足而提供一种基于复值词向量的神经网络模型的文本情感或所属类别分类的方法，搭建一个基于大量文本语料集，分别构建文本的词向量和相对位置信息，运用复值神经网络模型训练文本分类模型，并利用反向传播、随机下降法Adam训练网络模型得到最优模型在测试集上预测结果，最终得到更加准确的分类结果。

[0007] 本发明的目的是通过以下技术方案来实现的：

[0008] 一种基于位置与语义的复值词向量构建方法，包括如下步骤：

[0009] (1)采用jieba分词工具，对篇章和句子进行分词，从而构建多模态分类语料集

[0010] (2)从(1)构建的多模态分类语料集中,随机选取80%*N个样本作为训练集,10%*N个样本划分为验证集和剩余的10%*N个样本划分为测试集,并分别对训练集、验证集和测试集进行预处理;

[0011] (3)选取多模态分类语料集中预处理之后的句子构建复值神经网络模型,进而建立复值神经网络模型的损失函数:

[0012] $$loss = \sum_{i=1}^{N}[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)]$$

[0013] 其中:y_i代表真是类别标签,$\hat{y}_i$代表预测结果;

[0014] (4)将复值神经网络模型在训练集上进行训练获得语义分类模型;

[0015] (5)在验证集对训练后神经网络模型进行效果验证,并记录保存达到最优时的模型参数;

[0016] (6)用上一步中保存的最优的模型参数去测试测试集上的样本,最终得到每个测试样本的预测结果,对比测试标签,计算出分类准确率。

[0017] 所述(3)中复值神经网络模型构建,包括如下步骤:

[0018] 3.1、根据单词在篇章和句子的位置构造每个词的绝对位置索引,即:

[0019] $p_i = e^{i\beta POS}$

[0020] $$\beta = \frac{2\pi}{p_{j,i}}$$

[0021] 其中p_i表示当前词的位置向量,POS为其绝对位置,β为其初始化的周期,i为复数的虚部表示;

[0022] 3.2、运用glove工具得到每个篇章或句子中单词的300维词向量w_i,同时建立一个位置信息的矩阵,每一个位置索引对应一个300维的位置向量p_i进而构造每个词的相对位置索引,即:

[0023] $x_i = w_i e^{i\beta POS}$

[0024] 3.3、文本的每个词的词向量连同它的相对位置信息向量按其在句子中的顺序输入到Fasttext和LSTM,CNN网络中,具体计算公式如下:

[0025] Fasttext:

[0026] $Z^C = \sigma(Ax - By + b_r) + i\sigma(Bx - Ay + b_i)$

[0027] CNN

[0028] $Z^C = Ax - By + i(Bx + Ay)$

[0029] 其中,σ表示sigmod激活函数,A和B分别表示权重的实部和虚部,x和y表示输入特征的实部和虚部;

[0030] RNN

[0031] $$h_t^C = f(W^C h_{t-1}^C + V^C x_t^C + b^C)$$

[0032] 其中,x_t表示每个词的输入,$\vec{h}_t$表示分别得到的网络隐层状态;

[0033] 3.4、将上述步骤中最终的网络隐层输出$\vec{h}_t$输入到一个非线性全连接神经网络层得到神经网络表示向量x,再将表示x输入到softmax分类层输出最终的类别y:

[0034] $$\boldsymbol{x} = tanh(\mathbf{W}_R\vec{\boldsymbol{h}}_t + \boldsymbol{b}_R)$$

[0035] $y=\mathrm{softmax}(W_s x+b_s)$

[0036] 所述步骤3.2中每个词的相对位置索引是通过如下公式获得篇章或句子的向量表示：

[0037] 单词维度间相对位置：

[0038] $$x_{(j,\mathrm{pos+n})}=w_{j,k}e^{\frac{2\pi(pos+n)}{p_{j,k}}i}$$
$$=w_{j,k}e^{\frac{2\pi(pos)}{p_{j,k}}i}\times e^{\frac{2\pi n}{p_{j,k}}i}$$
$$=x(j,pos)\times e^{\frac{2\pi n}{p_{j,k}}i}$$

[0039] 其中：x表示单词的向量表示，$w_{j,k}$表示单词的语义向量，n表示单词间隔，pos，j，k的定义如上；

[0040] 单词间相对位置：

[0041] $$x_{(j_1,\mathrm{pos+n})}=w_{j_1,k}e^{\frac{2\pi(pos+n)}{p_{\cdot,k}}i}$$
$$=w_{j_1,k}e^{\frac{2\pi(pos)}{p_{\cdot,k}}i}\times\frac{w_{j_2,k}}{w_{j_1,k}}e^{\frac{2\pi n}{p_{\cdot,k}}i}$$
$$=x(j_2,pos)\times\frac{w_{j_2,k}}{w_{j_1,k}}e^{\frac{2\pi n}{p_{\cdot,k}}i}$$

[0042] 其中：p，k表示不同位置单词的第k个维度。

[0043] 本发明的有益效果是：

[0044] (1)搭建一个有效的实体文本语料集，克服了当前文本分类语料集匮乏的困境；

[0045] (2)利用文本或句子中单词的相对位置信息为特征，发展出一套基于位置信息的复值神经网络模型框架，进行分类任务。

附图说明

[0046] 图1为本发明的方法流程图；

[0047] 图2为3维复数词向量分布图；

[0048] 图3为基于不同词向量的分类模型运行时间对比。

[0049] 图4为不同词向量在相同句子下不同单词的相似度对比图。

具体实施方式

[0050] 下面结合附图进一步详细描述本发明的技术方案，但本发明的保护范围不局限于以下所述。图1显示了本方法提出的基于位置与语义的复值向量表示的神经网络分类方法的流程；图2显示了3维复数词向量的可能分布图；图3显示了不同词向量的分类模型运行时间对比结果；图4为不同词向量在相同句子下不同单词的相似度对比图。：

[0051] 传统的包含关键词的网页文字内容的提取，整理和加工完全靠手工，费事费力，且随着网页数据的爆炸型增长，单纯靠手工的方法变得效率低下，在这种情况下，本系统使用

网络爬虫获取网页内容，再经过数据处理形成有效的关键词网页语料库。

[0052] 基于已获取的网页数据，建立多模态数据集，具体步骤如下：

[0053] (1)：用jieba分词工具，对篇章和句子进行分词，并去除分词后的停止词，以及无用标点符号等，从而构建多模态分类语料集。所述多模态分类语料集为一个文本分类语料集，该语料集的总样本数为N，其中每条样本包含一段文本；

[0054] (2)：从(1)中构建的多模态分类语料集中，随机选取80％的文本作为训练集，10％的文本划分为验证集集，剩下的10％文本划分为测试集，并分别对训练集、验证集和测试集预处理；

[0055] (3)：对预处理之后的篇章或句子，根据单词在篇章或句子的位置构造中其绝对位置信息特征，并分别输入到复值Fasttext和LSTM，CNN神经网络模型中，运用方法如下：

[0056] 3.1：根据单词的位置构造篇章或句子每个词的绝对位置索引，假设一个词出现在句首，那么它的位置索引将被标记为1，周期为$\frac{2\pi}{p_{j,i}}$。而句子中的其他词的位置索引将依次叠加：

[0057] $p_i = e^{i\beta pos}$

[0058] $$\beta = \frac{2\pi}{p_{j,i}}$$

[0059] 其中p_i表示当前词的位置向量，POS为其绝对位置，β为其初始化的周期，i为复数的虚部表示。

[0060] 3.2：运用glove工具得到每个篇章或句子中单词的300维词向量w_i，同时建立一个位置信息的矩阵，每一个位置索引对应一个300维的位置向量p_i，模型初始化阶段用均匀分布初始化该参数矩阵，并在模型训练过程中更新优化。在这一步我们得到篇章和句子中的每个词的表示$x_i = w_i e^{i\beta POS}$。

[0061] 所述步骤3.2中每个词的相对位置索引是通过如下公式获得篇章或句子的向量表示：

[0062] 单词维度间相对位置推导如下：

[0063] $$p_{(j,pos+n)} = r_{j,k} e^{\frac{2\pi(pos+n)}{p_{j,k}} i} = r_{j,k} e^{\frac{2\pi(pos)}{p_{j,k}} i} \times e^{\frac{2\pi n}{p_{j,k}} i} = p(j,pos) \times e^{\frac{2\pi n}{p_{j,k}} i}$$

[0064] 单词间相对位置推导如下：

[0065] $$p_{(j_1,pos+n)} = r_{j_1,k} e^{\frac{2\pi(pos+n)}{p_{\cdot,k}} i} = r_{j_1,k} e^{\frac{2\pi(pos)}{p_{\cdot,k}} i} \times \frac{r_{j_2,k}}{r_{j_1,k}} e^{\frac{2\pi n}{p_{\cdot,k}} i} = p(j_2,pos) \times \frac{r_{j_2,k}}{r_{j_1,k}} e^{\frac{2\pi n}{p_{\cdot,k}} i}$$

[0066] 3.3：把文本的每个词的词向量连同它的位置信息向量按其在句子中的顺序输入

到Fasttext和LSTM,CNN网络中,具体计算公式如下:

[0067] Fasttext:

[0068] $Z^C=\sigma(Ax-By+b_r)+i\sigma(Bx-Ay+b_i)$

[0069] CNN:

[0070] $Z^C=Ax-By+i(Bx+Ay)$

[0071] RNN:

[0072] $$h_t^C = f(W^C h_{t-1}^C + V^C x_t^C + b^C)$$

[0073] 其中σ表示sigmod激活函数,A和B分别表示权重的实部和虚部,x和y表示输入特征的实部和虚部,RNN公式中权重与输入,输出均使用了复值表示。通过用上述的公式,输入每个词的表示300维x_t,分别得到网络隐层状态表示128维$\vec{h}_t$。

[0074] 3.4:根据上一步中得到的文本中每个词的网络隐层状态表示$\vec{h}_t$,并将其输入到一个非线性全连接神经网络层,从而得到神经网络表示向量x,再将表示x输入到softmax分类层输出最终的类别y:

[0075] $$\boldsymbol{x} = tanh(\mathbf{W}_R \vec{\boldsymbol{h}}_t + \boldsymbol{b}_R)$$

[0076] $y=softmax(W_s x+b_s)$

[0077] 定义复值神经网络模型损失函数为:

[0078] $$loss = \sum_{i=1}^{N} [y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i)]$$

[0079] 其中y_i代表真是类别标签,$\hat{y}_i$代表预测结果。通过反向传播算法、批量随机梯度下降法训练模型。通过反向传播算法、批量(mini-batch=32)Adam梯度下降法训练模型。

[0080] 在训练集上训练模型,每间隔100个批次,在验证集上进行验证模型效果,记录保存在验证集上效果达到最优时的模型参数。

[0081] 用上一步中保存的最优的模型去测试测试集上的样本,最终得到每个测试样本的预测结果,对比测试标签,计算出分类准确率。最终得到每个样本的预测结果,对比测试标签,计算出分类准确率,对比Fasttext模型、卷积神经网络模型、LSTM模型相较于复值模型,统计出其运行时间表格,可以非常直观的观察到本发明可以明显的提升分析模型的效果,如图3所示。为进一步证明模型方法的有效性,我们从数据集中随机挑选了一个句子,分别计算不同3-gram之间的相似度得分,从图4我们可以看出,当我们直接使用word-embedding时得出的打分在good except和except good间值很大,即颜色偏绿;使用transformer位置构建方法,得出的打分值基本固定,即位置向量的权重很大,这严重影响了,最后的结果(虽然具有规律性);最后图4(c)使用我们的复值词向量得出的打分值很小,颜色偏黄,原因是我们的词向量可以smooth词向量和位置向量,从而减少在网络中训练的参数。

[0082] 本发明方案所公开的技术手段不仅限于上述实施方式所公开的技术手段,还包括由以上技术特征任意组合所组成的技术方案。应当指出,对于本技术领域的普通技术人员来说,在不脱离本发明原理的前提下,还可以做出若干改进和润饰,这些改进和润饰也视为本发明的保护范围。

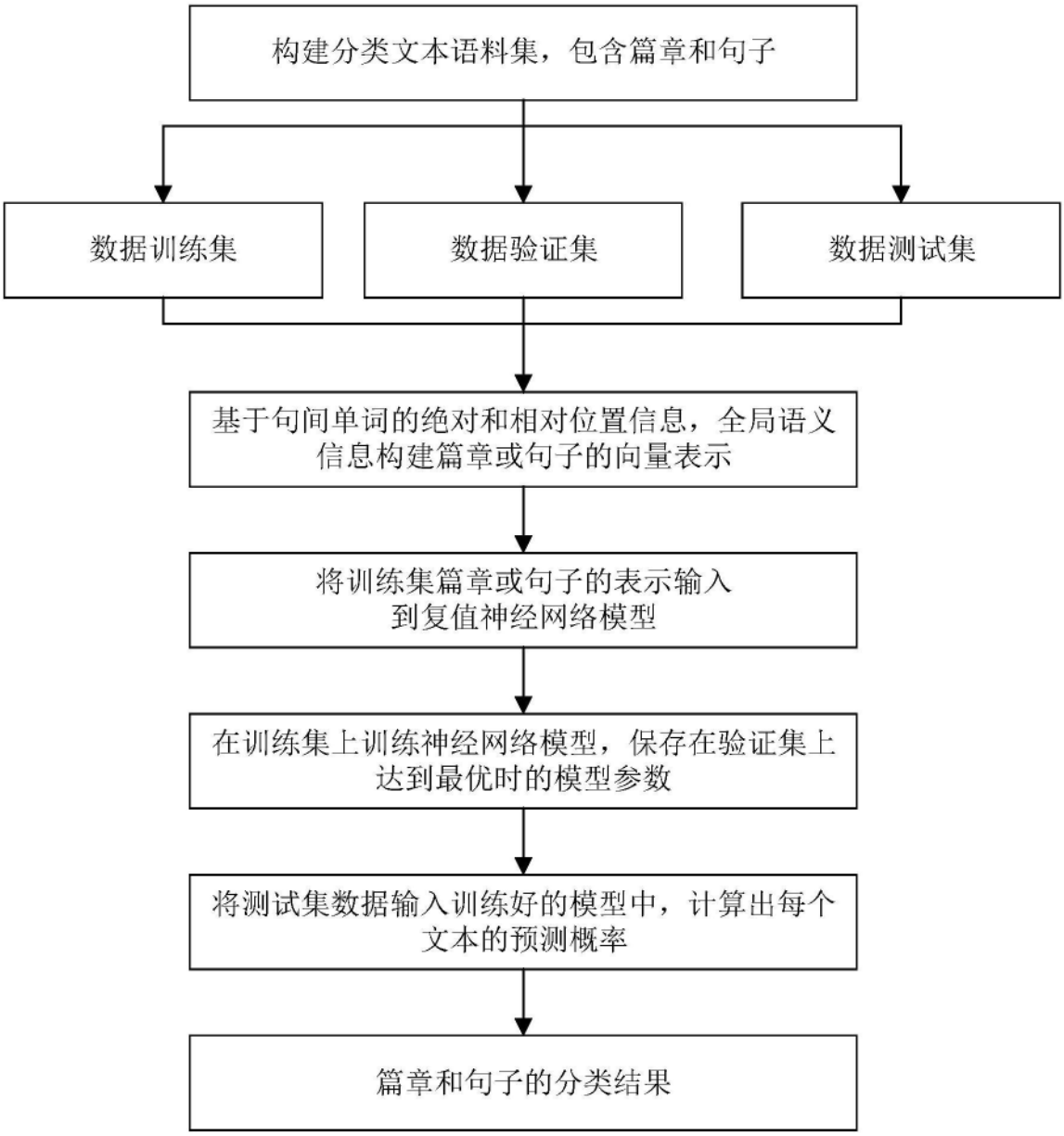

构建分类文本语料集，包含篇章和句子
数据训练集
数据验证集
数据测试集
基于句间单词的绝对和相对位置信息，全局语义信息构建篇章或句子的向量表示
将训练集篇章或句子的表示输入到复值神经网络模型
在训练集上训练神经网络模型，保存在验证集上达到最优时的模型参数
将测试集数据输入训练好的模型中，计算出每个文本的预测概率
篇章和句子的分类结果

图1

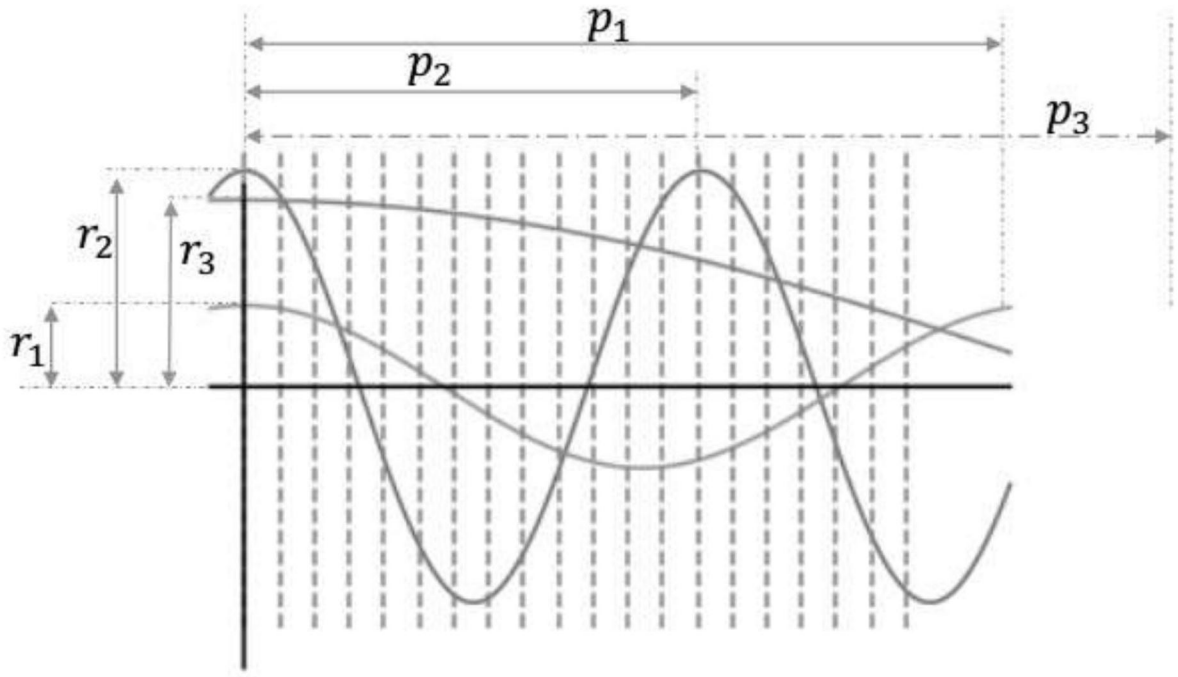

p_1
p_2
p_3
r_2
r_3
r_1

图2

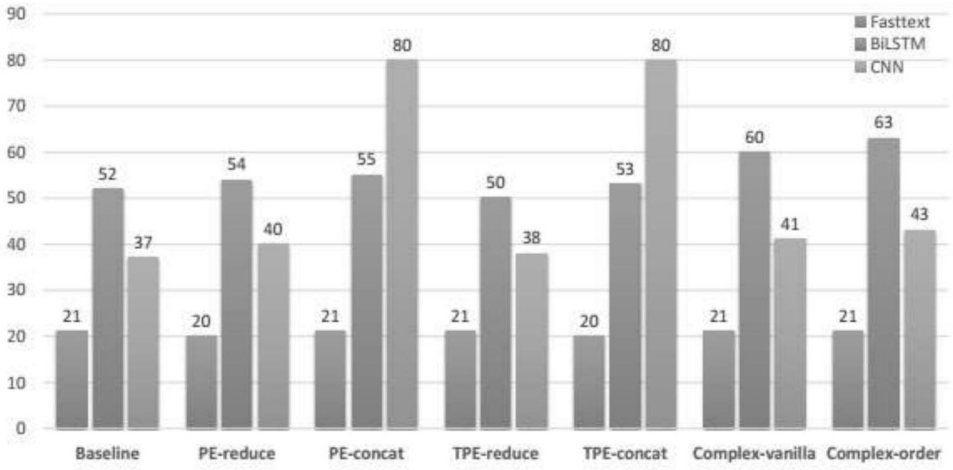

Fasttext
BiLSTM
CNN
90
80
70
60
50
40
30
20
10
0
21
52
37
20
54
40
21
55
80
21
50
38
20
53
80
21
60
41
21
63
43
Baseline
PE-reduce
PE-concat
TPE-reduce
TPE-concat
Complex-vanilla
Complex-order

图3

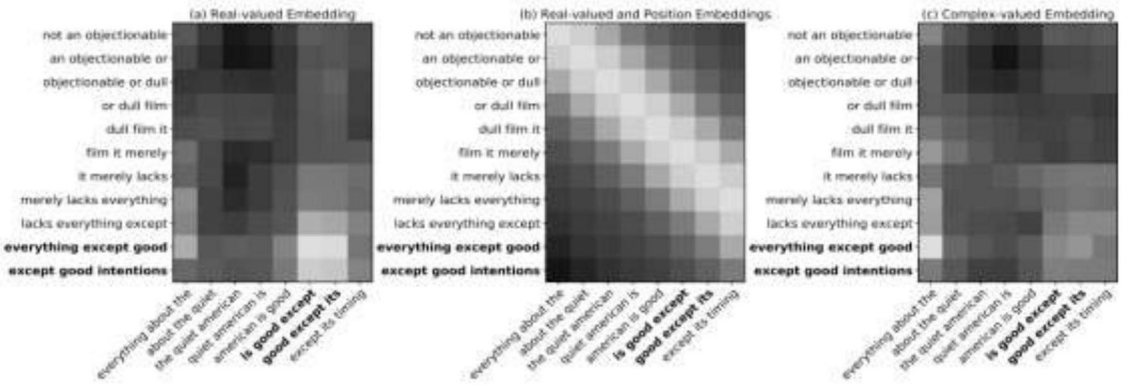

(a) Real-valued Embedding
(b) Real-valued and Position Embeddings
(c) Complex-valued Embedding
not an objectionable
an objectionable or
objectionable or dull
or dull film
dull film it
film it merely
it merely lacks
merely lacks everything
lacks everything except
everything except good
except good intentions
everything about the
about the quiet
the quiet american
quiet american is
american is good
is good except
good except its
except its timing

图4