(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 110569355 B
(45)授权公告日 2022.05.03

(21)申请号 201910671527.X
(22)申请日 2019.07.24
(65)同一申请的已公布的文献号
申请公布号 CN 110569355 A
(43)申请公布日 2019.12.13
(73)专利权人 中国科学院信息工程研究所
地址 100093 北京市海淀区闵庄路甲89号
(72)发明人 虎嵩林 周艳 朱福庆 韩冀中
(74)专利代理机构 北京君尚知识产权代理有限公司 11200
代理人 邵可声
(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/33(2019.01)
G06F 40/289(2020.01)

(56)对比文件
CN 108470061 A,2018.08.31
CN 104731770 A,2015.06.24
US 2012148161 A1,2012.06.14
Peng Chen,Zhongqian Sun, Lidong.Recurrent attention network on memory for aspect sentiment analysis.《EMNLP2017》.2017,
Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, Songlin Hu.Conditional BERT Contextual Augmentation.《ICCS 2019》.2019,

审查员 徐捷

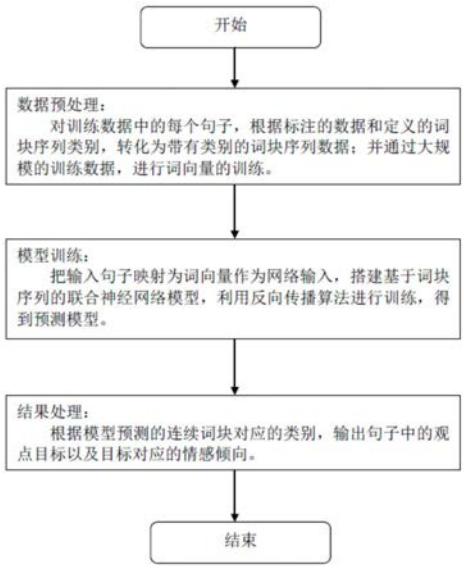
权利要求书3页 说明书7页 附图2页

(54)发明名称

一种基于词块的观点目标抽取和目标情感分类联合方法及系统

(57)摘要

本发明提出一种基于词块的观点目标抽取和目标情感分类联合方法及系统,具体为:对于每个连续词块,设计词块级别的特征以此来充分利用多个词之间的整体信息;计算每个词块的情感信息而非单独计算每一个词的情感信息,这样保证词块里多个词的情感倾向的一致性。本发明一是通过有效利用多个词整体信息,二是通过为多个词组成的词块计算一个情感信息表示来避免情感不一致的问题,来提升抽取和分类的准确率,具有良好的实用性。



CN 110569355 B

1.一种基于词块的观点目标抽取和目标情感分类联合方法,包括如下步骤:

把需要进行观点目标抽取和目标情感分类的句子进行处理,得到每个句子中所有的连续词块;

把得到的句子信息以及词块信息输入到观点目标抽取和目标情感分类的联合模型中,得到句子中每一词块对应的各类别概率分布;

根据每一词块对应的各类别概率分布,获取句子中的观点目标及其对应的情感类别;

其中,所述观点目标抽取和目标情感分类的联合模型的构建方法,包括:

(1)获取训练数据输入句子中每一个词的词类别标签与每一词块的词块类别标签,并将数量数据输入句子及相应词块信息输入联合模型;

(2)利用通过word2vec训练得到的词向量矩阵与预训练好的语言模型,分别将每个词映射成传统词向量与基于上下文的词向量,并进行向量拼接,以生成每个词对应的词向量;

(3)基于每个词对应的词向量、词类别标签及词块信息,联合模型输出各词块对应的各类别概率分布;

(4)基于各类别概率分布与相应的词块类别标签计算交叉熵损失函数的值来进行反向梯度传播,获取模型参数,以构建观点目标抽取和目标情感分类的联合模型。

2.如权利要求1所述一种基于词块的观点目标抽取和目标情感分类联合方法,其特征在于,联合模型的构建方法步骤(1)具体方法包括:

(1-1)设定词块的最大长度值,枚举训练数据输入句子中不超过所设定的最大长度的所有连续文本词块;所述最大长度值N的设定范围为1≤N≤L,其中,N为整数,L为输入句子的最大长度;

(1-2)定义词块的类别集合为4个类别{TPOS,TNEG,TNEU,O},其中,TPOS表示词块是观点目标且其情感倾向是积极的,TNEG表示词块是观点目标且其情感倾向是消极的,TNEU表示词块是观点目标且其情感倾向是中性的,O表示词块不是情感目标;根据标注语料中句子标注的观点目标以及目标情感分类,为所有的词块标记类别。

3.如权利要求2所述一种基于词块的观点目标抽取和目标情感分类联合方法,其特征在于,训练数据输入句子$X=\{w_1,w_2,\ldots,w_t\ldots,w_T\}$,其中$w_t$表示输入句子中的第t个词;模型的目标是预测词块集合中每个词块的类别$Y=\{(i,J,l)\mid 1\leqslant i\leqslant j\leqslant T;j-i+1\leqslant L;l\in C\}$,其中i,j表示词块在句子中的起始位置和终止位置,l表示词块对应的标签,C表示词块的类别集合。

4.如权利要求3所述一种基于词块的观点目标抽取和目标情感分类联合方法,其特征在于,通过以下步骤获取各词块对应的各类别概率分布:

3-1)在上下文表示层,把句子中每个词对应的词向量作为输入,采用多层双向长短记忆神经网络学习句子中每个词的上下文表示向量;

3-2)每个词块采用两种词块级别的信息来对其进行表示:一种是词块的边界信息,一种是词块的整体信息;

3-3)基于词块的注意力机制用来计算输入句子中和每个词块相关联的上下文中的情感信息;

3-4)在输出层把每个词块的向量表示以及基于词块注意力机制的上下文情感信息表示拼接在一起,获取各词块对应的各类别概率分布。

5.如权利要求4所述一种基于词块的观点目标抽取和目标情感分类联合方法，其特征在于，步骤3-1)中所述的上下文表示为：

$$\left\{h_1^{(m)},\dots h_t^{(m)},\dots,h_T^{(m)}\right\}=BiLSTM^{(m)}\{h_1^{(m-1)},\dots h_t^{(m-1)},\dots,h_T^{(m-1)}\}$$

其中，$h_t^{(m)}$表示第m层中第t个隐藏单元的状态，第M层的隐藏层状态$\left\{h_1^{(M)},\dots h_t^{(M)},\dots,h_T^{(M)}\right\}$作为每个词的上下文表示，其中m∈{1，…，M}，M为多层双向长短记忆神经网络的层数。

6.如权利要求4所述一种基于词块的观点目标抽取和目标情感分类联合方法，其特征在于，步骤3-2)中所述边界信息用边界词对应的stack BiLSTM层的输出来进行表示；所述整体信息采用词块中所有词的上下文信息和进行表示；任意一个词块(i，j)词块表示为：

$$s_{(i,j)}=[h_i^{(M)};\sum_{k=i}^{j}h_k^{(M)};h_j^{(M)}]$$

其中，$h_i^{(M)}$，$h_k^{(M)}$，$h_j^{(M)}$是多层双向长短记忆神经网络的输出，这两类信息的表示向量拼接起来来作为词块的表示。

7.如权利要求4所述一种基于词块的观点目标抽取和目标情感分类联合方法，其特征在于，步骤3-3)具体方法包括：

3-3-1)每个词和词块(i，j)的距离，来定义这个词的权重w′t：

$$w_t'=1-\frac{l_t}{T}$$

其中l_t表示第t个词到词块(i，j)的距离；对于词块中的词，设置距离l_t的值为0；对于词块左边的词，距离l_t为到词块最左边词的距离；对于词块右边的词，距离l_t为到词块最右边词的距离；

3-3-2)根据上面获取的权重值，模型计算每个词块(i，j)基于位置权重的上下文表示：

$$e_t=w_t'*\ h_t^{(M)}$$

其中e_t为第t个词对应的基于距离权重的表示，w'_t表示第t个词的权重值，$h_t^{(M)}$表示第t个词的上下文；

3-3-3)利用基于词块的注意力机制来计算和词块相关的上下文信息，包括：

a)对于词块(i，j)，计算每个上下文词e_t和它的相关性权重，计算公式为：

$$f(s(i,j),e_t)=\tanh(s_{(i,j)}W_\alpha\ e_t^T+\ b_\alpha)$$

$$\alpha_{(i,j)}^t=\frac{\exp(f(s_{(i,j)},e_t))}{\sum_{t=1}^{T}\exp(f(s_{(i,j)},e_t))}$$

其中$s_{(i,j)}$为词块(i，j)的向量表示，e_t为第t个词对应的基于距离权重的表示，W_α为权重矩阵，b_α为偏移向量，$\alpha_{(i,j)}^t$为第t个词和词块(i，j)的相关程度向量；

b)把这些权重和这些词上下文表示向量相乘并进行加权,即可得到和词块(i,j)相关的上下文信息表示$c_{(i,j)}$:

$$c_{(i,j)} = \sum_{t=1}^{T} \alpha_{(i,j)}^{t} e_t。$$

8.如权利要求4所述一种基于词块的观点目标抽取和目标情感分类联合方法,其特征在于,步骤3-4)中把上述得到的词块(i,j)的向量表示以及与它相关的上下文表示拼接在一起,用来预测词块(i,j)的类别:

$r_{(i,j)}=[s_{(i,j)};c_{(i,j)}]$

$y=\mathrm{softmax}(W_y r_{(i,j)}+b_y)$

其中y为词块对应到各类别的概率分布,W_y为权重矩阵,b_y为偏移向量。

9.如权利要求1所述一种基于词块的观点目标抽取和目标情感分类联合方法,其特征在于,步骤(4)中对所有训练样本,通过最大化样本的最大似然函数来训练模型,更新模型中的参数,训练的目标函数loss定义如下:

$$\mathrm{loss} = -\sum_{h\in H}\sum_{i\in S_h} g_i \log(y_i)$$

其中,g_i是词块对应的真实分类的向量表示,y_i为预测得到的概率分布,H代表所有的训练的句子,S_h表示第h个句子中所有的词块。

10.一种基于词块的观点目标抽取和目标情感分类联合系统,包括:

待测数据预处理模块:把需要进行观点目标抽取和目标情感分类的句子进行处理,得到每个句子中所有的连续词块;

词块分类预测模块:把得到的句子信息以及词块信息输入到观点目标抽取和目标情感分类的联合模型中,得到句子中每一词块对应的各类别概率分布;

结果获取模块:根据每一词块对应的各类别概率分布,获取句子中的观点目标及其对应的情感类别;

其中,所述观点目标抽取和目标情感分类的联合模型的构建方法,包括:

(1)获取训练数据输入句子中每一个词的词类别标签与每一词块的词块类别标签,并将数量数据输入句子及相应词块信息输入联合模型

(2)利用通过word2vec训练得到的词向量矩阵与预训练好的语言模型,分别将每个词映射成传统词向量与基于上下文的词向量,并进行向量拼接,以生成每个词对应的词向量;

(3)基于每个词对应的词向量、词类别标签及词块信息,联合模型输出各词块对应的各类别概率分布;

(4)基于各类别概率分布与相应的词块类别标签计算交叉熵损失函数的值来进行反向梯度传播,获取模型参数,以构建观点目标抽取和目标情感分类的联合模型。

一种基于词块的观点目标抽取和目标情感分类联合方法及系统

技术领域：

[0001] 本发明涉及深度学习与自然语言处理技术，具体涉及一种基于词块的观点目标抽取和目标情感分类联合方法及系统。

背景技术：

[0002] 近年来，互联网信息技术高速发展，新闻、社交等网站每天有海量的新数据产生出来，这些数据中包含着各种各样表达观点或者情感的信息。对这些数据进行观点、立场、态度等的分析，可以帮助人们更好的做出判断以及决策，比如：对商品的评论进行分析，可以了解用户对商品的满意度，从而制定更加合理的营销策略。但是由于互联网上的数据量以几何倍数增长，如何从这些海量的信息中查找出对自己有用的数据并进行正确的分析，已经成为了一项非常用意义的研究课题。

[0003] 情感分析技术就是一项针对用户产生的信息进行情感倾向进行分析研究的技术。根据情感分析的粒度主要分为：篇章级情感分析，句子级情感分析以及目标级别情感分析。其中针对目标的情感分析主要包括两个子任务，一个是找出句子中的观点目标，另外一个是判断对该目标的情感倾向。传统的基于目标的情感分析分别研究其中一个子任务，但是在实际应用中往往不仅仅需要完成其中一个子任务，而是既需要抽取出其中的观点目标同时又要对目标的情感倾向进行分类。一种比较直观的做法是把两个子任务以流水线的方式串联起来执行，但是这样无法利用两个子任务之间的相互联系。为了充分利用这两个子任务之间的联系，一些基于词级别序列标注的联合方法被提了出来。很多观点目标是由多个词组成的而非单个词，比如“鼠标左键”是由“鼠标”和“左键”两个词构成，所以基于词级别序列标注的联合方法处理这类观点目标的时候仍然存在一些局限性，一是很难利用观点目标级别的特征，二是预测出的同一个观点目标的多个词之间的情感倾向有可能存在不一致，比如对于“鼠标左键”这个情感目标，可能对“鼠标”这个词给出的标签情感分类是正向的，但是“左键”这个词给出的情感分类是负向的。

发明内容：

[0004] 针对上述技术问题，本发明提出一种基于词块的观点目标抽取和目标情感分类联合方法及系统，来利用情感目标级别的特征同时避免情感分类不一致的问题。

[0005] 为了解决上述技术问题，本发明的技术方案如下：

[0006] 一种基于词块的观点目标抽取和目标情感分类联合方法，包括如下步骤：

[0007] 把需要进行观点目标抽取和目标情感分类的句子进行处理，得到每个句子中所有的连续词块；

[0008] 把得到的句子信息以及词块信息输入到观点目标抽取和目标情感分类的联合模型中，对词块进行分类预测；

[0009] 根据词块的类别获取句子中的观点目标及其对应的情感类别；其中，所述观点目

标抽取和目标情感分类的联合模型的构建方法,包括:

[0010] (1)将训练数据中句子以及其中标注的观点目标以及目标情感转化为对应的词块以及词块对应的类别;同时通过大规模的非标注语料,训练得到具有语义信息的词向量;

[0011] (2)将训练数据的句子中的每个词映射成对应的词向量,输入基于词块的目标抽取和目标情感分类联合神经网络模型,并通过反向传播算法进行训练;

[0012] (3)将需进行观点目标抽取和目标情感分类的句子输入训练完成的联合预测模型,预测出每个词块对应的类别,根据词块的类别得到句子中的观点目标以及对该目标的情感类别。

[0013] 进一步地,所述联合模型的构建方法步骤(1)具体方法包括:

[0014] (1-1)设定词块的最大长度值,枚举训练数据输入句子中不超过所设定的最大长度的所有连续文本词块;

[0015] (1-2)根据标注语料中句子标注的观点目标以及目标情感分类,为所有的词块标记类别;

[0016] (1-3)利用传统的词向量以及基于上下文的词向量来表示输入句子中的每一个词:通过word2vec在大规模的非标注语料上训练获取到传统词向量,把句子输入到预训练好的ELMo(Embeddings from Language Models)模型中得到基于上下文的词向量。

[0017] 更进一步地,步骤(1-1)中最大长度值N的设定范围为$1\leqslant N\leqslant L$,其中,N为整数,L为输入句子的最大长度,优选的长度为4。

[0018] 更进一步地,步骤(1-2)中定义词块的类别集合为4个类别{TPOS,TNEG,TNEU,O},这4个类别代表的含义分别为:TPOS表示词块是观点目标且其情感倾向是积极的,TNEG表示词块是观点目标且其情感倾向是消极的,TNEU表示词块是观点目标且其情感倾向是中性的,O表示词块不是情感目标。

[0019] 进一步地,所述联合模型的构建方法步骤(2)中所述模型输入为:

[0020] 包含T个词的句子$X=\{w_1,w_2,\cdots,w_T\}$,其中w_t表示输入句子中的第t个词,模型的目标是预测词块集合中每个词块的类别$Y=\{(i,j,l)\mid 1\leqslant i\leqslant j\leqslant T;j-i+1\leqslant L;l\in C\}$,其中i,j表示词块在句子中的起始位置和终止位置,l表示词块对应的标签,C表示类别集合。

[0021] 进一步地,所述联合模型的构建方法步骤(2)中所述模型训练过程包括:

[0022] 2-1)将上述输入句子传统的词向量以及基于上下文的词向量进行拼接,作为下一层的输入;

[0023] 2-2)在上下文表示层,把句子中每个词对应的词向量作为输入,采用多层双向长短记忆神经网络(stacked Bi-LSTM)学习句子中每个词的上下文表示向量;

[0024] 2-3)每个词块采用两种词块级别的信息来对其进行表示:一种是词块的边界信息,一种是词块的整体信息;

[0025] 2-4)基于词块的注意力机制用来计算输入句子中和每个词块相关联的上下文中的情感信息;

[0026] 2-5)在输出层把每个词块的向量表示以及基于词块注意力机制的上下文情感信息表示拼接在一起,用于预测词块的类别;

[0027] 2-6)选取交叉熵为模型训练的损失函数;

[0028] 2-7)通过反向传播算法训练模型,更新模型中所有的参数,最终得到词块分类模

型。

[0029] 更进一步地，步骤2-2)中所述的上下文表示为：

[0030] $$\left\{h_1^{(m)},\dots h_t^{(m)},\dots,h_T^{(m)}\right\}=BiLSTM^{(m)}\{h_1^{(m-1)},\dots h_t^{(m-1)},\dots,h_T^{(m-1)}\}$$

[0031] 其中，$h_t^{(m)}$表示第m层t个隐藏单元的状态，第M层的隐藏层状态$\left\{h_1^{(M)},\dots h_t^{(M)},\dots,h_T^{(M)}\right\}$作为每个词的上下文表示。

[0032] 更进一步地，步骤2-3)中所述边界信息用边界词对应的stack BiLSTM层的输出来进行表示；所述整体信息采用词块中所有词的上下文信息和进行表示；任意一个词块(i,j)词块表示为：

[0033] $$s_{(i,j)}=[h_i^{(M)};\sum_{k=i}^{j}h_k^{(M)};h_j^{(M)}]$$

[0034] 其中，$h_i^{(M)}$，$h_k^{(M)}$，$h_j^{(M)}$是多层双向长短记忆神经网络的输出，这两类信息的表示向量拼接起来来作为词块的表示。

[0035] 更进一步地，步骤2-4)中由于任务不仅需要识别出词块是否是观点目标，还需要判断出这个词块对应的情感信息，而这些情感信息往往是在上下文中，所以采用基于连续词块的注意力机制来计算文本中和目标相关的情感信息。直观上来说，离一个连续词块越近的词，对这个连续的词块的影响可能越大，采用基于距离权重的上下文信息来表示这种影响，对于离词块越近的词设置的权重越大，离其越远的词设置的权重越小。

[0036] 步骤2-4)具体方法包括：

[0037] 2-4-1)每个词和词块(i,j)的距离，来定义这个词的权重w'_t：

[0038] $$w'_t=1-\frac{l_t}{T}$$

[0039] 其中l_t表示第t个词到词块(i,j)的距离；对于词块中的词，设置距离l_t的值为0；对于词块左边的词，距离l_t为到词块最左边词的距离；对于词块右边的词，距离l_t为到词块最右边词的距离；

[0040] 2-4-2)根据上面获取的权重值，模型计算每个词块(i,j)基于位置权重的上下文表示：

[0041] $$e_t=w'_t*h_t^{(M)}$$

[0042] 其中e_t为第t个词对应的基于距离权重的表示，w'_t表示第t个词的权重值，$h_t^{(M)}$表示第t个词的上下文；

[0043] 2-4-3)利用基于词块的注意力机制来计算和词块相关的上下文信息，包括：

[0044] a)对于词块(i,j)，计算每个上下文词e_t和它的相关性权重，计算公式为：

[0045] $$f(s(i,j),e_t)=\tanh(s_{(i,j)}W_a\,e_t^T+b_a)$$

[0046] $$\alpha_{(i,j)}^t=\frac{\exp(f(s_{(i,j)},e_t))}{\sum_{k=1}^{T}\exp(f(s_{(i,j)},e_k))}$$

[0047] 其中$s_{(i,j)}$为词块(i,j)的向量表示，e_t为第t个词对应的基于距离权重的表示，W_a为权重矩阵，b_a为偏移向量，$\alpha_{(i,j)}^t$为第t个词和词块(i,j)的相关相关程度向量；

[0048] b)把这些权重和这些词上下文表示向量相乘并进行加权，即可得到和词块(i,j)相关的上下文信息表示$c_{(i,j)}$：

[0049] $$c_{(i,j)} = \sum_{t=1}^{T} \alpha_{(i,j)}^t e_t$$

[0050] 更进一步地，步骤2-5)中把上述得到的词块(i,j)的向量表示以及与它相关的上下文表示拼接在一起，用来预测词块(i,j)的类别：

[0051] $r_{(i,j)} = [s_{(i,j)}; c_{(i,j)}]$

[0052] $y = \mathrm{softmax}(W_y r_{(i,j)} + b_y)$

[0053] 其中y为词块对应到各类别的概率分布，W_y为权重矩阵，b_y为偏移向量。

[0054] 对词块的分类进行处理，在类别O里的认为不是观点目标，在类别TPOS里的即为句子中情感倾向为正向的观点目标，在类别TNEG中的即为句子中情感倾向为负向的观点目标，在类别TNEU中的即为句子中情感倾向为中性的观点目标。

[0055] 更进一步地，步骤2-6)中对所有训练样本，通过最大化样本的最大似然函数来训练模型，更新模型中的参数，训练的目标函数loss定义如下：

[0056] $$\mathrm{loss} = -\sum_{h \in H} \sum_{i \in S_h} g_i \log(y_i)$$

[0057] 其中g_i是词块对应的真实分类的向量表示，y_i为预测得到的概率分布，H代表所有的训练的句子，S_h表示第h个句子中所有的词块。

[0058] 一种基于词块的观点目标抽取和目标情感分类联合系统，包括：

[0059] 待测数据预处理模块：把需要进行观点目标抽取和目标情感分类的句子进行处理，得到每个句子中所有的连续词块；

[0060] 词块分类预测模块：把得到的句子信息以及词块信息输入到观点目标抽取和目标情感分类的联合模型中，对词块进行分类预测；所述联合模型通过把输入句子映射为词向量作为网络输入，搭建基于词块序列的联合神经网络模型，利用反向传播算法进行训练所得；

[0061] 结果获取模块：根据词块的类别获取句子中的观点目标及其对应的情感类别。

[0062] 本发明的有益效果在于：针对基于词级别序列标注的观点目标抽取和目标情感倾向分类方法中的两个问题：一是，很难利用多个词组成的观点目标整体信息，二是，多个词组成的观点目标之间可能存在情感不一致；提出了基于连续词块的联合模型来同时进行观点目标抽取以及目标情感倾向分类，具体为：对于每个连续词块，设计词块级别的特征以此来充分利用多个词之间的整体信息；计算每个词块的情感信息而非单独计算每一个词的情感信息，这样保证词块里多个词的情感倾向的一致性。这样，本发明一是通过有效利用多个词整体信息，二是通过为多个词组成的词块计算一个情感信息表示来避免情感不一致的问题，来提升抽取和分类的准确率，具有良好的实用性。

附图说明：

[0063] 图1为本发明实施例提供的基于词块级别的观点目标抽取和目标情感分类流程图；

[0064] 图2为本发明实施例的神经网络模型结构图。

具体实施方式：

[0065] 为使本发明的上述目的、特征和优点能够更加明显易懂，下面通过具体实施案例并结合附图，对本发明做进一步详细说明。

[0066] 图1为本实施例中基于词块级别的观点目标抽取和目标情感分类流程图方法的流程图，如图所示，该方法主要包括三个阶段，分别是：数据预处理阶段，基于词块级别的观点目标抽取和目标情感分类联合模型训练阶段，对预测得到的分类进行匹配获取到句子中的观点目标以及其情感倾向类别阶段。

[0067] (一)数据预处理阶段

[0068] 步骤1对于每个句子，穷举其中所有的连续词块(词块长度设定上限)。根据标注语料中给出的观点目标和目标情感分类数据，得到所有词块所在的分类。分类的类别包含4种：TPOS表示词块是观点目标且对该目标的情感类别是正向的，TNEG表示词块是观点目标且对该目标的情感类别是负向的，TNEU表示词块是观点目标且对该目标的情感类别是中性的，O表示词块不是观点目标。比如对于句子“硬盘坏了”，长度上限设置为2，则所有的连续词块及其对应的标签为“(硬,O)，(盘,O)，(坏,O)，(了,O)，(硬盘,TNEG)，(盘坏,O)，(坏了,O)”。

[0069] 步骤2，用无标注的语料，通过word2vec训练得到具有语义信息的词向量表示，提供给模型使用。

[0070] (二)模型训练阶段

[0071] 结合图2，基于词块级别的观点目标抽取和目标情感分类联合模型包括以下具体步骤：

[0072] 步骤1，形式化输出和输出，输入为包含T个词的句子$X=\{w_1,w_2,\cdots,w_T\}$，其中w_t表示输入句子中的第t个词，目标是预测词块集合中每个词块的类别$Y=\{(i,j,l)\mid 1\leqslant i\leqslant j\leqslant T;j-i+1\leqslant L;l\in C\}$，其中i，j表示词块在句子中的起始位置和终止位置，l表示词块对应的标签，C表示类别集合；

[0073] 步骤2，利用通过word2vec训练得到的词向量表示以及预训练好的语言模型，分别将输入句子中的每个词映射成对应的词向量两类词向量，并且把这两种类别的词向量进行拼接；

[0074] 步骤3，上下文表示层，把句子中每个词对应的词向量作为输入，采用M层的双向长短记忆神经网络(stackedBi-LSTM)学习输入句子中每个词的上下文信息，其中第m($m\in\{1,\cdots,M\}$)层的隐藏层状态计算公式如下：

[0075] $$\left\{h_1^{(m)},\dots h_t^{(m)},\dots,h_T^{(m)}\right\}=\mathrm{BiLSTM}^{(m)}\{h_1^{(m-1)},\dots h_t^{(m-1)},\dots,h_T^{(m-1)}\}$$

[0076] 其中，$h_t^{(m)}$表示第m层t个隐藏单元的状态。把第M层的隐藏层状态

$\left\{h_1^{(M)},\dots h_t^{(M)},\dots,h_T^{(M)}\right\}$作为每个词的上下文表示；

[0077] 步骤3词块的表示层，对于任意一个词块(i，j)，用两种词块级别的信息来对其进行表示：一种是词块的边界信息，一种是词块的整体信息。其中边界信息直接用上一层得到的边界词的向量表示来获取，整体信息则是上一层得到的词块中所有词的表示的加和：

[0078] $$s_{(i,j)}=[h_i^{(M)};\sum_{k=i}^{j}h_k^{(M)};h_j^{(M)}]$$

[0079] 其中，$h_i^{(M)}$，$h_k^{(M)}$，$h_j^{(M)}$是多层双向长短记忆神经网络的输出。把这两类信息的表示向量拼接起来来作为词块的表示；

[0080] 步骤4基于词块的注意力机制，由于模型不仅需要预测一个词块是否是情感目标还需要预测句中对于该目标的情感倾向，而这些情感信息往往存在于词块的上下文中，为了获取和词块相关的情感信息，提出了基于词块的注意力机制来学习这些相关信息；

[0081] 步骤4-1，直观上来说，离一个词块越近的上下文词，对该词块的影响越大。模型来用了基于位置权重的上下文来模拟这种影响。首先根据每个词和词块(i，j)的距离，来定义这个词的权重w'_t：

[0082] $$w'_t=1-\frac{l_t}{T}$$

[0083] 其中l_t表示第t个词到词块(i，j)的距离。对于词块中的词，设置距离l_t的值为0；对于词块左边的词，距离l_t为到词块最左边词的距离；对于词块右边的词，距离l_t为到词块最右边词的距离。

[0084] 步骤4-2根据上面获取的权重值，模型计算每个词块(i，j)基于位置权重的上下文表示：

[0085] $$e_t=w'_t*h_t^{(M)}$$

[0086] 其中e_t为第t个词对应的基于距离权重的表示，w'_t表示第t个词的权重值，$h_t^{(M)}$表示第t个词的上下文。

[0087] 步骤4-3获取到基于位置权重的上下文表示之后，利用基于词块的注意力机制来计算和词块相关的上下文信息。首先对于词块(i，j)，计算每个上下文词e_t和它的相关性权重，计算公式为：

[0088] $$f(s(i,j),e_t)=\tanh(s_{(i,j)}W_a e_t^T+b_a)$$

[0089] $$\alpha_{(i,j)}^t=\frac{\exp(f(s_{(i,j)},e_t))}{\sum_{k=1}^{T}\exp(f(s_{(i,j)},e_k))}$$

[0090] 其中$s_{(i,j)}$为词块(i，j)的向量表示，e_t为第t个词对应的基于距离权重的表示，W_a为权重矩阵，b_a为偏移向量，$\alpha_{(i,j)}^t$为第t个词和词块(i，j)的相关相关程度向量。把这些权重和这些词上下文表示向量相乘并进行加权，即可得到和词块(i，j)相关的上下文信息表示$c_{(i,j)}$：

[0091] $$c_{(i,j)} = \sum_{t=1}^{T} \alpha_{(i,j)}^{t} e_t$$

[0092] 步骤5,输出层,把上面得到的词块(i,j)的向量表示以及与它相关的上下文表示拼接在一起,用来预测词块(i,j)的类别:

[0093] $r_{(i,j)} = [s_{(i,j)}; c_{(i,j)}]$

[0094] $y = \mathrm{softmax}(W_y r_{(i,j)} + b_y)$

[0095] 其中y为词块对应到各类别的概率分布,W_y为权重矩阵,b_y为偏移向量。

[0096] 步骤6,对所有训练样本,通过最大化样本的最大似然函数来训练模型,更新模型中的参数,训练的目标函数loss定义如下:

[0097] $$loss = -\sum_{h \in H} \sum_{i \in S_h} g_i \log(y_i)$$

[0098] 其中g_i是词块对应的真实分类的向量表示,y_i为预测得到的概率分布,H代表所有的训练的句子,S_h表示第h个句子中所有的词块。

[0099] (三)结果处理阶段

[0100] 步骤1,把需要进行观点目标抽取和目标情感分类的句子进行处理,得到每个句子中所有的连续词块;

[0101] 步骤2,把得到的句子信息以及词块信息输入到上面得到的观点目标抽取和目标情感分类的联合模型中,对词块进行分类预测;

[0102] 步骤3,根据词块的类别获取到句子中的观点目标以及其对应的情感类别

[0103] 由上述方案可以看出,本方案针对基于单个词级别序列标注的联合目标抽取和目标情感分类模型中两个问题:一时无法利用组成目标的多个词的整体信息,二是多个词之间可能存在情感分类不一致的问题,提出基于词块的联合抽取和分类模型,可以提高模型预测的性能,具有良好的实用性。

[0104] 以上所述仅是本发明的优选实施方式,应当指出,对于本技术领域的普通技术人员,在不脱离本发明构思的前提下,还可以做出若干改进和润饰,这些改进和润饰也应视为本发明保护范围内。

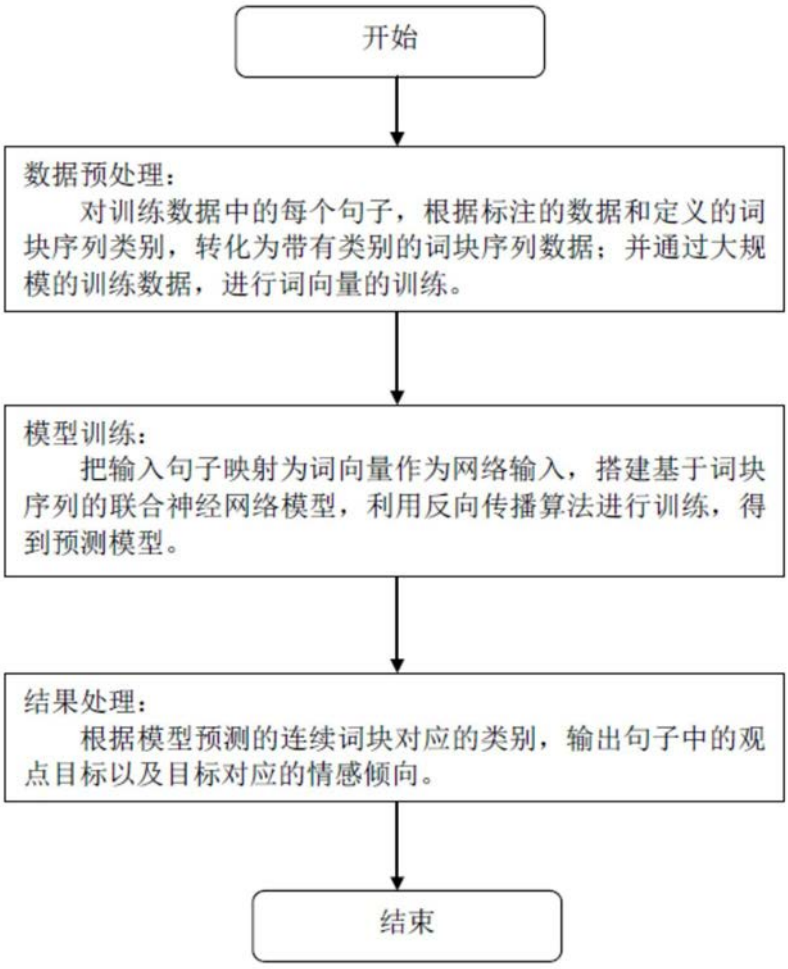

开始
数据预处理：
对训练数据中的每个句子，根据标注的数据和定义的词块序列类别，转化为带有类别的词块序列数据；并通过大规模的训练数据，进行词向量的训练。
模型训练：
把输入句子映射为词向量作为网络输入，搭建基于词块序列的联合神经网络模型，利用反向传播算法进行训练，得到预测模型。
结果处理：
根据模型预测的连续词块对应的类别，输出句子中的观点目标以及目标对应的情感倾向。
结束

图1

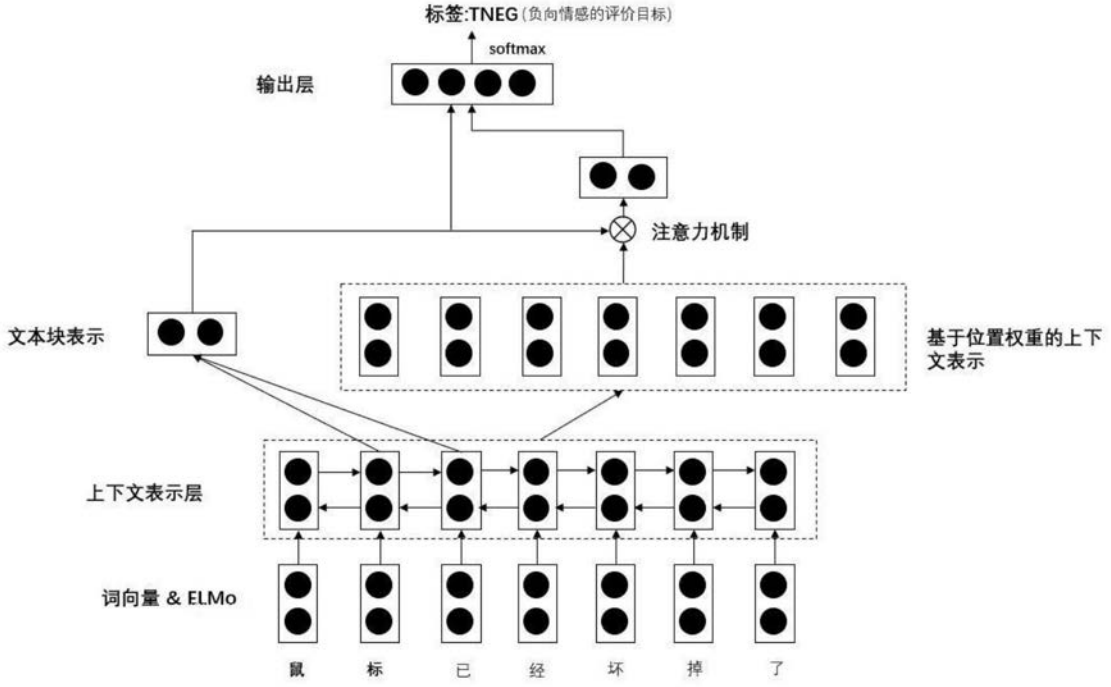

标签:TNEG(负向情感的评价目标)
softmax
输出层
注意力机制
文本块表示
基于位置权重的上下文表示
上下文表示层
词向量 & ELMo
鼠
标
已
经
坏
掉
了

图2