(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114942976 A
(43)申请公布日 2022.08.26

(21)申请号 202210599136.3

(22)申请日 2022.05.30

(71)申请人 北京邮电大学
地址 100876 北京市海淀区西土城路10号

(72)发明人 李睿凡 翟泽鹏 冯方向 张光卫 王小捷

(74)专利代理机构 北京挺立专利事务所(普通合伙) 11265
专利代理师 高福勇

(51)Int.Cl.
G06F 16/33(2019.01)
G06F 16/35(2019.01)
G06K 9/62(2022.01)

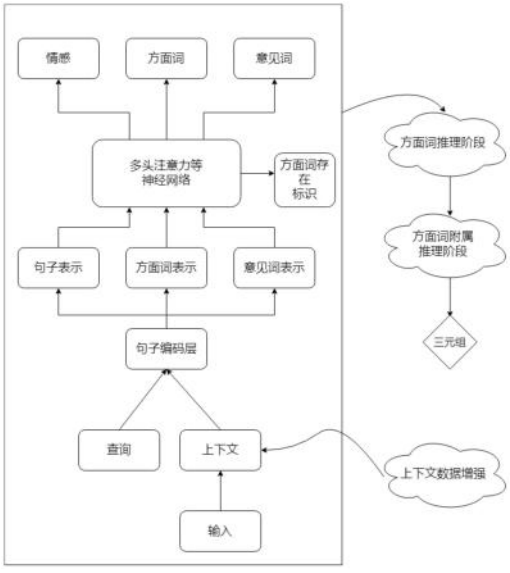
权利要求书3页 说明书9页 附图4页

(54)发明名称

基于掩码上下文机器阅读理解的方面情感三元组抽取方法

(57)摘要

本发明公开了一种基于掩码上下文机器阅读理解的方面情感三元组抽取方法，在推理时，应用先推理方面词再掩码无关方面词推理意见词，可以有效减少其他方面词干扰问题；在训练时，应用上下文数据增强，有效地扩充了语料并为推理打下基础；在模型结构方面，设计了四个模块协同工作，这四个模块包括方面词提取模块、意见词提取模块、情感分类模块以及方面词存在探测模块，通过以上三个要素，有效解决了以往MRC方法面临的方面词干扰问题。



CN 114942976 A

1.一种基于掩码上下文机器阅读理解的方面情感三元组抽取方法,其特征在于:

在方面词推理阶段,使用BERT作为句子的编码器,输入一个固定的查询q和一个原始句子作为上下文,经过模型得到方面词a以及方面词存在标识e,若标识结果为True,则将得到的方面词a加入到方面词集合A中,将上下文把集合A中所有方面词掩码作为掩码上下文,与查询q再次输入至模型中得到方面词a以及方面词存在标识e,重复此流程,直到标识结果为False,得到方面词集合A;探测经过掩码之后的上下文是否仍存在方面词,如果所有的方面词均被掩码,其标识为False,否则为True,得到句子表示、方面词表示和意见词表示,再通过多头注意力神经网络融合句子表示、方面词表示和意见词表示的信息,得到情感s,输出情感s、方面词a、意见词o和方面词存在标识e;

在方面词附属推理阶段,对于方面词集合A中的每个方面词a,在上下文中直接掩码掉除了方面词a以外所有无关的方面词,根据查询q以及掩码所有无关方面词的上下文得到方面词a对应的意见词O集合以及情感s,最后输出句子存在的所有的方面情感三元组(a,o,s)。

2.根据权利要求1所述的基于掩码上下文机器阅读理解的方面情感三元组抽取方法,其特征在于,模型包括方面词提取模块、意见词提取模块、情感分类模块以及方面词存在探测模块,模型的处理流程为:将固定的查询与掩码上下文输入至BERT中,然后将掩码上下文对应的输出向量输入至上下文表示层得到掩码上下文的表示,方面词提取模块根据掩码上下文的表示通过方面词表示层得到方面词的表示,并通过方面词判定线性层得到方面词;意见词提取模块根据掩码上下文的表示通过意见词表示层以及意见词判定线性层得到意见词;情感分类模块利用多头注意力模块,将上下文表示、意见词与方面词分别作为查询、键、值,起到融合三者信息作用,并且通过层归一化、最大池化模块以及情感判定线性层得到情感;方面词存在探测模块则将BERT中[CLS]对应的表示向量、方面词表示的最大池化、意见词表示的最大池化拼接起来,再通过存在判定线性层得到方面词存在标识。

3.根据权利要求1所述的基于掩码上下文机器阅读理解的方面情感三元组抽取方法,其特征在于,掩码矩阵M如下:

$$M_{ij}=\begin{cases}-\infty, & if\ j=k\\ 0, & otherwise\end{cases} \qquad (1)$$

将掩码矩阵应用到注意力矩阵中:

$$A(Q,K,V)=softmax\left(\frac{QK^T}{\sqrt{d}}+M\right)V \qquad (2)$$

其中,Q、K、V分别代表注意力中的查询、键以及值,d是对应维度。

4.根据权利要求1所述的基于掩码上下文机器阅读理解的方面情感三元组抽取方法,其特征在于,固定的查询q设置为:查询q中第一个方面词以及对应的意见词。

5.根据权利要求1所述的基于掩码上下文机器阅读理解的方面情感三元组抽取方法,其特征在于,将固定的查询q以及来自上下文数据增强中的X作为BERT的输入,输入形式为[CLS]q[SEP]X[SEP],假定q包含m个单词,X包含n个单词,d是对应维度,从最后一个BERT层得到的表示记为$h\in\mathbb{R}^{d\times(m+n+3)}$,上下文表示以及token[CLS]的表示分别记为$h_x\in\mathbb{R}^{d\times n}$和$h_{cls}\in\mathbb{R}^{d\times 1}$。

6.根据权利要求5所述的基于掩码上下文机器阅读理解的方面情感三元组抽取方法，其特征在于，使用公式(4)获取方面词表示r_a：

$$r_a = W_{a,1} h_x \quad (4)$$

使用公式(5)和公式(6)分别计算方面词的开始与结束位置的概率$p^{a,s}$和$p^{a,e}$：

$$p^{a,e} = \mathrm{softmax}(W_{a,2} r_a) \quad (5)$$

$$p^{a,e} = \mathrm{softmax}(W_{a,3} r_a) \quad (6)$$

其中，$W_{a,1} \in \mathbb{R}^{d \times d}, W_{a,2} \in \mathbb{R}^{1 \times d}, W_{a,3} \in \mathbb{R}^{1 \times d}$是可训练的权重；

使用公式(7)计算方面词的开始或结束位置的损失$\mathcal{L}_A$：

$$\mathcal{L}_A = -\sum_{i=1}^{n} y_i^{a,s} log\, p_i^{a,s} - \sum_{i=1}^{n} y_i^{a,e} log\, p_i^{a,e} \quad (7)$$

其中$y^{a,s}$和$y^{a,e} \in \mathbb{R}^n$分别是对于句子中首个未掩码的方面词开始与结束位置的真实值。

7.根据权利要求6所述的基于掩码上下文机器阅读理解的方面情感三元组抽取方法，其特征在于，使用公式(8)获取意见词表示r_o：

$$r_o = W_{o,1} h_x \quad (8)$$

使用公式(9)和公式(10)分别计算意见词的开始与结束位置的概率$p^{o,s}$和$p^{o,e}$：

$$p^{o,s} = \mathrm{softmax}(W_{o,2} r_o) \quad (9)$$

$$p^{o,e} = \mathrm{softmax}(W_{o,3} r_o) \quad (10)$$

其中，$W_{o,1} \in \mathbb{R}^{d \times d}, W_{o,2} \in \mathbb{R}^{1 \times d}, W_{o,3} \in \mathbb{R}^{1 \times d}$是可训练的权重；

使用公式(11)计算意见词的开始或结束位置的损失$\mathcal{L}_O$：

$$\mathcal{L}_O = -\sum_{i=1}^{n} y_i^{o,s} log\, p_i^{o,s} - \sum_{i=1}^{n} y_i^{o,e} log\, p_i^{o,e} \quad (11)$$

其中，$y^{o,s}$和$y^{o,e} \in \mathbb{R}^n$分别是对于句子中首个未掩码的方面词对应的意见词开始与结束位置的真实值。

8.根据权利要求7所述的基于掩码上下文机器阅读理解的方面情感三元组抽取方法，其特征在于，使用多头自注意力机制融合方面词、意见词以及上下文的信息，使用公式(12)获取情感表示r_s：

$$r_s = \mathrm{LN}(h_x + \mathrm{MultiHead}(h_x, r_a, r_o)) \quad (12)$$

其中，LN、MultiHead分别代表层归一化模块、多头注意力网络，其参数中查询、键、值分别为h_x、r_a、r_o；

使用公式(13)计算g_s：

$$g_s = \mathrm{MP}(r_s) \quad (13)$$

其中，$\mathbf{g_s} \in \mathbb{R}^{d \times n}$代表方面词、意见词以及上下文信息的融合表示，MP代表最大池化；

使用公式(14)计算p^s：

$$p^s = \mathrm{softmax}(W_s g_s + b_s) \quad (14)$$

其中，p^s代表情感概率，$\mathbf{W_s} \in \mathbb{R}^{1 \times d}$为可训练的权重，$b_s$为偏置项；

使用公式(15)表达情感分类损失：

$$\mathcal{L}_S = -\sum_{i=1}^{3} y_i^s \log p_i^s \quad (15)$$

其中，$y^s \in \mathbb{R}^3$是真实的情感极性标签。

9.根据权利要求8所述的基于掩码上下文机器阅读理解的方面情感三元组抽取方法，其特征在于，使用公式(16)获取r_e：

$$r_e = h_{cls} \oplus MP(r_o) \oplus MP(r_a) \quad (16)$$

其中，$r_e \in \mathbb{R}^{3d\times 1}$是中间变量，代表方面词存在探测表示，$\oplus$表示拼接操作；

使用公式(17)计算p^e：

$$p^e = \mathrm{softmax}(W_e r_e + b_e) \quad (17)$$

其中，p^e代表方面词的存在概率，$W_e \in \mathbb{R}^{2\times 3d}$和$b_e$分别是可训练的权重和偏置项；

使用公式(18)计算二元交叉熵损失$\mathcal{L}_E$：

$$\mathcal{L}_E = -\sum_{i=1}^{2} y_i^e \log p_i^e \quad (18)$$

其中，$y^e \in \mathbb{R}^2$为方面词存在真实标签，p^e为预测的方面词存在概率。

10.根据权利要求9所述的基于掩码上下文机器阅读理解的方面情感三元组抽取方法，其特征在于，根据句子中方面词的数量指数级扩增数据集，考虑上下文信息预测方面词、意见词以及情感设计模型，模型的目标损失函数表示为：

$$\mathcal{L}_T = \alpha\mathcal{L}_A + \beta\mathcal{L}_O + \gamma\mathcal{L}_S + \delta\mathcal{L}_E \quad (19)$$

其中，α，β，γ以及δ分别是调整$\mathcal{L}_A$、$\mathcal{L}_O$、$\mathcal{L}_S$、$\mathcal{L}_E$损失影响因子的超参数。

基于掩码上下文机器阅读理解的方面情感三元组抽取方法

技术领域

[0001] 本发明涉及自然语言处理技术领域,尤其涉及一种基于掩码上下文机器阅读理解的方面情感三元组抽取方法。

背景技术

[0002] 方面级情感分析(Aspect-based Sentiment Analysis,ABSA)是一个细粒度情感分析任务,其旨在提取意见以及情感信息。ABSA通常包含三个子任务:方面词抽取(Aspect Term Extraction,ATE)、意见词抽取(Opinion Term Extraction,OTE)以及方面级情感分析(Aspect Sentiment Classification,ASC)。

[0003] 方面三元组抽取(Aspect Sentiment Triplet Extraction,ASTE)是ABSA的新型子任务,该任务目标是为了提取句子中蕴含的方面词、意见词、情感三元组。如图1所示,给定句子为“Nice ambience,but highly overrated place”,其三元组为(ambience,Nice,positive)以及(place,overrated,negative)。

[0004] 为解决ASTE任务,早期的方法应用了两阶段的管道式架构,其首先分别鉴别方面词和意见词并将他们配对加上情感极性形成三元组,然而这种方法忽略了三元组之间信息的交互,会潜在地导致错误传播问题。为解决这该问题,近期研究用一种端到端的方法联合提取三元组,但这些方法的表现仍不令人满意,尤其是遇到一句话包含多个方面词时,其中一个原因是缺乏专门针对多方面词的设计。

[0005] 最近形成了一种有效的基于多轮机器阅读理解(Machine Reading Comprehension,MRC)的方法应用于ASTE任务,诸如实体关系抽取、命名实体识别以及事件抽取。对于ASTE任务来说,基于MRC的方法是设置不同的查询(query)且使用相同的上下文(context)并通过多轮问答(QA)解决。MRC包含两个阶段,第一阶段是方面词推理(Aspect Inference,AI)阶段,目的是构造一个关于方面词的查询去提取出所有的方面词,例如“What aspects?”;第二阶段是方面词附属推理(Aspect Accessory Inference,AAI)阶段,目的是构造一个关于意见词和情感的查询去提取出所有的意见词和情感,例如“What opinions and sentiment given the aspect ambience?”。

[0006] 尽管之前基于MRC的方法取得了很好的效果,然而当一句话包含多个方面词时,传统的MRC方法在AAI阶段会存在其他方面词带来干扰的严重问题,进而导致方法失效。图1展示了一个方面词带来干扰的例子:在“ambience”的方面词附属推理阶段,应用于MRC的注意力(attention)机制容易受到另外一个方面词“place”的干扰,反之亦然。需要注意的是,目前包含多个方面词的句子占整个数据集接近一半。因此,如何有效地解决多个方面词带来的干扰问题,对于提升MRC方法在ASTE任务上的性能至关重要。

发明内容

[0007] 本发明的目的是解决ASTE任务中包含多个方面词的句子的干扰问题,提出一种基于掩码上下文机器阅读理解(Context-Masked machine reading comprehension,COM-

MRC)的方面情感三元组抽取方法,来提升了MRC方法在ASTE任务上的性能。

[0008] 为了实现上述目的,本发明提供如下技术方案:

[0009] 一种基于掩码上下文机器阅读理解的方面情感三元组抽取方法,包括以下步骤:

[0010] 在方面词推理阶段,使用BERT(Bidirectional Encoder Representation from Transformers)作为句子的编码器,输入一个固定的查询q和一个原始句子作为上下文,经过模型得到方面词a以及方面词存在标识e,若标识结果为True,则将得到的方面词a加入到方面词集合A中,将上下文把集合A中所有方面词掩码作为掩码上下文,与查询q再次输入至模型中得到方面词以及方面词存在标识e,重复此流程,直到标识结果为False,得到方面词集合A;探测经过掩码之后的上下文是否仍存在方面词,如果所有的方面词均被掩码,其标识为False,否则为True,得到句子表示、方面词表示和意见词表示,再通过多头注意力神经网络融合句子表示、方面词表示和意见词表示的信息,得到情感s,输出情感s、方面词a、意见词o和方面词存在标识e;

[0011] 在方面词附属推理阶段,对于方面词集合A中的每个方面词a,在上下文中直接掩码掉除了方面词a以外所有无关的方面词,根据查询q以及掩码所有无关方面词的上下文得到方面词a对应的意见词O集合以及情感s,输出句子存在的所有的方面情感三元组(a,o,s)。

[0012] 进一步地,模型包括方面词提取模块、意见词提取模块、情感分类模块以及方面词存在探测模块,模型的处理流程为:将固定的查询与掩码上下文输入至BERT中,然后将掩码上下文对应的输出向量输入至上下文表示层得到掩码上下文的表示,方面词提取模块根据掩码上下文的表示通过方面词表示层得到方面词的表示,并通过方面词判定线性层得到方面词;意见词提取模块根据掩码上下文的表示通过意见词表示层以及意见词判定线性层得到意见词;情感分类模块利用多头注意力模块,将上下文表示、意见词与方面词分别作为查询、键、值,起到融合三者信息作用,并且通过层归一化、最大池化模块以及情感判定线性层得到情感;方面词存在探测模块则将BERT中[CLS]对应的表示向量、方面词表示的最大池化、意见词表示的最大池化拼接起来,再通过存在判定线性层得到方面词存在标识。

[0013] 进一步地,对于一个句子S=$\{w_1,w_2,...,w_n\}$,其长度为n,假设句子有t个方面词,在每个方面词上应用两种操作:掩码或者不掩码,在训练中一条训练数据扩充为2^t条训练数据,掩码的第k个单词设置其注意力分值为0。

[0014] 进一步地,掩码矩阵M如下:

[0015] $$M_{ij}=\begin{cases}-\infty, & if\ j=k\\ 0, & otherwise\end{cases} \quad (1)$$

[0016] 将掩码矩阵应用到注意力矩阵中:

[0017] $$A(Q,K,V)=softmax\left(\frac{QK^T}{\sqrt{d}}+M\right)V \quad (2)$$

[0018] 进一步地,固定的查询q设置为:查询q中第一个方面词以及对应的意见词。

[0019] 进一步地,将固定的查询q以及来自上下文数据增强中的X作为BERT的输入,输入形式为[CLS]q[SEP]X[SEP],假定q包含m个单词,X包含n个单词,d是对应维度,从最后一个BERT层得到的表示记为$h\in\mathbb{R}^{d\times(m+n+3)}$,上下文表示以及token[CLS]的表示分别记为

$h_x \in \mathbb{R}^{d\times n}$和$h_{cls} \in \mathbb{R}^{d\times 1}$。

[0020] 进一步地，使用公式(4)获取方面词表示r_a：

[0021] $r_a = W_{a,1} h_x$ (4)

[0022] 使用公式(5)和公式(6)分别计算方面词的开始与结束位置的概率$p^{a,s}$和$p^{a,e}$：

[0023] $p^{a,s} = \mathrm{softmax}(W_{a,2} r_a)$ (5)

[0024] $p^{a,e} = \mathrm{softmax}(W_{a,3} r_a)$ (6)

[0025] 其中，$W_{a,1} \in \mathbb{R}^{d\times d}, W_{a,2} \in \mathbb{R}^{1\times d}, W_{a,3} \in \mathbb{R}^{1\times d}$是可训练的权重；

[0026] 使用公式(7)计算方面词的开始或结束位置的损失$\mathcal{L}_A$：

[0027] $$\mathcal{L}_A = -\sum_{i=1}^{n} y_i^{a,s} \log p_i^{a,s} - \sum_{i=1}^{n} y_i^{a,e} \log p_i^{a,e} \quad (7)$$

[0028] 其中$y^{a,s}$和$y^{a,e} \in \mathbb{R}^n$分别是对于句子中首个未掩码的方面词开始与结束位置的真实值。

[0029] 进一步地，使用公式(8)获取意见词表示ro：

[0030] $r_o = W_{o,1} h_x$ (8)

[0031] 使用公式(9)和公式(10)分别计算意见词的开始与结束位置的概率$p^{o,s}$和$p^{o,e}$：

[0032] $p^{o,s} = \mathrm{softmax}(W_{o,2} r_o)$ (9)

[0033] $p^{o,e} = \mathrm{softmax}(W_{o,3} r_o)$ (10)

[0034] 其中，$W_{o,1} \in \mathbb{R}^{d\times d}, W_{o,2} \in \mathbb{R}^{1\times d}, W_{o,3} \in \mathbb{R}^{1\times d}$是可训练的权重；

[0035] 使用公式(11)计算意见词的开始或结束位置的损失$\mathcal{L}_O$：

[0036] $$\mathcal{L}_O = -\sum_{i=1}^{n} y_i^{o,s} \log p_i^{o,s} - \sum_{i=1}^{n} y_i^{o,e} \log p_i^{o,e} \quad (11)$$

[0037] 其中，$y^{o,s}$和$y^{o,e} \in \mathbb{R}^n$分别是对于句子中首个未掩码的方面词对应的意见词开始与结束位置的真实值。

[0038] 进一步地，使用多头自注意力机制融合方面词、意见词以及上下文的信息，使用公式(12)获取情感表示r_s：

[0039] $r_s = \mathrm{LN}(h_x + \mathrm{MultiHead}(h_x, r_a, r_o))$ (12)

[0040] 其中，LN、MultiHead分别代表层归一化(Layer Normalization)模块、多头注意力(Multihead Attention)网络，其参数中查询(query)、键(key)、值(value)分别为h_x、r_a、r_o；

[0041] 使用公式(13)计算g_s：

[0042] $g_s = \mathrm{MP}(r_s)$ (13)

[0043] 其中，$g_s \in \mathbb{R}^{d\times n}$代表方面词、意见词以及上下文信息的融合表示，MP代表最大池化(Max Pooling)；

[0044] 使用公式(14)计算p^s：

[0045] $p^s = \mathrm{softmax}(W_s g_s + b_s)$ (14)

[0046] 其中，p^s代表情感概率，$W_s \in \mathbb{R}^{1\times d}$为可训练的权重，$b_s$为偏置项；

[0047] 使用公式(15)表达情感分类损失：

[0048] $$\mathcal{L}_S = -\sum_{i=1}^{3} y_i^s \log p_i^s \quad (15)$$

[0049] 其中，$y^s \in \mathbb{R}^3$是真实的情感极性标签。

[0050] 进一步地，使用公式(16)获取r_e：

[0051] $$r_e = h_{cls} \oplus MP(r_o) \oplus MP(r_a) \quad (16)$$

[0052] 其中，$r_e \in \mathbb{R}^{3d\times 1}$是中间变量，代表方面词存在探测的表示，$\oplus$表示拼接操作；

[0053] 使用公式(17)计算p^e：

[0054] $$p^e = \mathrm{softmax}(W_e r_e + b_e) \quad (17)$$

[0055] 其中，p^e代表方面词的存在概率，$W_e \in \mathbb{R}^{2\times 3d}$和$b_e$分别是可训练的权重和偏置；

[0056] 使用公式(18)计算二元交叉熵损失$\mathcal{L}_E$：

[0057] $$\mathcal{L}_E = -\sum_{i=1}^{2} y_i^e \log p_i^e \quad (18)$$

[0058] 其中，$y^e \in \mathbb{R}^2$为方面词存在真实标签，p^e为预测的方面词存在概率。

[0059] 进一步地，根据句子中方面词的数量指数级扩增数据集，考虑上下文信息预测方面词、意见词以及情感设计模型，模型目标函数表示为：

[0060] $$\mathcal{L}_T = \alpha\mathcal{L}_A + \beta\mathcal{L}_O + \gamma\mathcal{L}_S + \delta\mathcal{L}_E \quad (19)$$

[0061] 其中，α，β，γ以及δ分别是调整$\mathcal{L}_A$、$\mathcal{L}_O$、$\mathcal{L}_S$、$\mathcal{L}_E$损失影响因子的超参数。

[0062] 与现有技术相比，本发明的有益效果为：

[0063] 本发明提出的基于掩码上下文机器阅读理解(COM-MRC)的方面情感三元组抽取方法，在推理时，应用先推理方面词再掩码无关方面词推理意见词，可以有效减少其他方面词干扰问题；在训练时，应用上下文数据增强，有效地扩充了语料并为推理打下基础；在模型结构方面，设计了四个模块协同工作，这四个模块包括方面词提取模块、意见词提取模块、情感分类模块以及方面词存在探测模块，通过以上三个要素，有效解决了以往MRC方法面临的方面词干扰问题。

附图说明

[0064] 为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明中记载的一些实施例，对于本领域普通技术人员来讲，还可以根据这些附图获得其他的附图。

[0065] 图1为传统MRC与本发明方法(COM-MRC)在ASTE任务中的主要不同之处。

[0066] 图2为本发明实施例提供的基于掩码上下文机器阅读理解(COM-MRC)的方面情感三元组抽取方法的模型框架图。

[0067] 图3为本发明实施例提供的基于掩码上下文机器阅读理解(COM-MRC)的方面情感三元组抽取方法的训练流程以及推理流程图。

[0068] 图4为本发明实施例提供的基于掩码上下文机器阅读理解(COM-MRC)的方面情感三元组抽取方法的流程图。

具体实施方式

[0069] 为了解决ASTE任务中包含多个方面词的句子的干扰问题,本发明提出一个有效的方法,即Context-Masked machine reading comprehension(COM-MRC)解决ASTE难题。我们的COM-MRC框架包含了三个要素:

[0070] 首先,本发明的推理算法会在方面词推理阶段提取所有的方面词,然后在方面词附属推理阶段我们并不是在查询中考虑当前处理的方面词,而是在上下文中直接掩码掉其他所有的方面词,由此模型将不会关注到其他的方面词从而减少他们带来的干扰。

[0071] 其次,为了更好地应用我们的推理策略,我们提出了一个新颖的上下文数据增强(context augmentation)方法。想法是为了通过多样的掩码上下文(masked context)增强模型的表现。在实践中,我们设计了一个常规的查询并配套不同的掩码上下文,因此该方法显著扩充了训练语料,也就是说当一个句子包含t个方面词时,训练语料扩充为2^t条训练样本。

[0072] 最后,为了适应推理算法,我们设计了一个有效的模型架构。共包含四个模块,分别是方面词推理模块、意见词推理模块、情感识别模块以及方面词存在判定模块,这四个模块协同工作。

[0073] 为了更好地理解本技术方案,下面结合附图对本发明的方法做详细的说明。

[0074] 问题定义如下:

[0075] 给定一个的句子S×{w_1,w_2,...,w_n},其长度为n,本发明方法模型的目标是输出该句中存在的所有的方面情感三元组(a,o,s),a和o分别表示方面词和意见词,方面词的情感极性s属于情感标签集合$\mathcal{S} = \{\mathrm{POS}, \mathrm{NEG}, \mathrm{NEU}\}$,情感标签集合由三种情感极性组成,分别是积极、消极和中性。

[0076] 本发明的基于掩码上下文机器阅读理解(COM-MRC)的方面情感三元组抽取方法,如图4所示:

[0077] 在方面词推理阶段,使用BERT作为句子的编码器,输入一个固定的查询q和一个原始句子作为上下文,经过模型得到方面词a以及方面词存在标识e,若标识结果为True,则将得到的方面词a加入到方面词集合A中,将上下文把集合A中所有方面词掩码作为掩码上下文,与查询q再次输入至模型中得到方面词a以及方面词存在标识e,重复此流程,直到标识结果为False,由此在方面词推理阶段我们得到了方面词集合A。

[0078] 探测经过掩码之后的上下文是否仍存在方面词,如果所有的方面词均被掩码,其标识为False,否则为True,得到句子表示、方面词表示和意见词表示,再通过多头注意力神经网络融合句子表示、方面词表示和意见词表示的信息,得到情感s,输出情感s、方面词a、意见词o和方面词存在标识e;

[0079] 在方面词附属推理阶段,对于方面词集合A中的每个方面词a,在上下文中直接掩码掉除了a以外所有无关的方面词,根据查询q以及掩码所有无关方面词的上下文得到方面词a对应的意见词O集合以及情感s。最后便可输出句子存在的所有的方面情感三元组(a,o,s)。

[0080] 关于输入:

[0081] 本发明使用BERT作为句子的编码器,输入包括一个固定的查询(query)和一个由上下文数据增强(Context Augmentation)得到的掩码上下文(masked context)。

[0082] 假设句子S有t个方面词,我们在每个方面词上应用两种操作:掩码或者不掩码。由此,在训练中一条训练数据便可扩充为2^t条训练数据,数据量扩充很多。在具体实现中,掩码第k个单词意为设置其注意力分值为0。具体来说,我们定义一个掩码矩阵M如下:

[0083] $$M_{ij}=\begin{cases}-\infty, & if\ j=k\\ 0, & otherwise\end{cases} \quad (1)$$

[0084] 然后将掩码矩阵应用到注意力矩阵中:

[0085] $$A(Q,K,V)=softmax\left(\frac{QK^T}{\sqrt{d}}+M\right)V \quad (2)$$

[0086] 其中,Q、K、V分别代表注意力中的查询(query)、键(key)以及值(value),d是对应维度。上下文数据增强以及训练细节如图3所示。

[0087] 本发明设计了一个固定的查询q去提示模型,提取最左侧(第一个)的方面词以及对应的意见词,查询设置如下:

[0088] q=Find the first aspect term and corresponding opinion terms in the text (3)

[0089] 将等式(3)中的查询q以及来自上下文数据增强中的X作为BERT的输入,输入形式为[CLS]q[SEP]X[SEP]。假定q包含m个单词且X包含n个单词,从最后一个BERT层得到的表示记为$\mathbf{h}\in\mathbb{R}^{d\times(m+n+3)}$,则上下文表示以及token[CLS]的表示分别记为$\mathbf{h_x}\in\mathbb{R}^{d\times n}$和$\mathbf{h_{cls}}\in\mathbb{R}^{d\times 1}$。

[0090] 关于本发明方法的模型架构,如图2所示,包括四个模块,分别是方面词提取模块、意见词提取模块、情感分类模块以及方面词存在探测模块。

[0091] 如图2所示,将固定的查询与掩码上下文输入至BERT中,然后将掩码上下文对应的输出向量输入至上下文表示层得到掩码上下文的表示。方面词提取模块根据掩码上下文的表示通过方面词表示层得到方面词的表示,并通过方面词判定线性层得到方面词,同理,意见词提取模块根据掩码上下文的表示通过意见词表示层以及意见词判定线性层得到意见词;情感分类模块利用多头注意力模块,将上下文表示、意见词与方面词分别作为查询、键、值,起到融合三者信息作用,并且通过层归一化、最大池化模块以及情感判定线性层得到情感;方面词存在探测模块则将BERT中[CLS]对应的表示向量、方面词表示的最大池化、意见词表示的最大池化拼接起来,再通过存在判定线性层得到方面词存在标识。

[0092] 1、关于方面词提取模块:

[0093] 基于区间方法,使用公式(4)获取方面词表示r_a:

[0094] $r_a=W_{a,1}h_x$ (4)

[0095] 使用公式(5)和公式(6)分别计算方面词的开始与结束位置的概率$p^{a,s}$和$p^{a,e}$:

[0096] $p^{a,s}=\mathrm{softmax}(W_{a,2}r_a)$ (5)

[0097] $p^{a,e}=\mathrm{softmax}(W_{a,3}r_a)$ (6)

[0098] 其中,$W_{a,1}\in\mathbb{R}^{d\times d}$,$W_{a,2}\in\mathbb{R}^{1\times d}$,$W_{a,3}\in\mathbb{R}^{1\times d}$是可训练的权重;

[0099] 使用公式(7)计算方面词的开始或结束位置的损失$\mathcal{L}_A$:

[0100] $$\mathcal{L}_A = -\sum_{i=1}^{n} y_i^{a,s}\log p_i^{a,s} - \sum_{i=1}^{n} y_i^{a,e}\log p_i^{a,e} \quad (7)$$

[0101] 其中$y^{a,s}$和$\mathbf{y}^{a,e} \in \mathbb{R}^n$分别是对于句子中首个未掩码的方面词开始与结束位置的真实值。

[0102] 2、关于意见词提取模块：

[0103] 与方面词提取模块类似，我们用同样的方法获得带意见词表示r_o以及损失$\mathcal{L}_O$，意见词表示r_o的获取公式为：

[0104] $$r_o = W_{o,1}h_x \quad (8)$$

[0105] 使用公式(9)和公式(10)分别计算意见词的开始与结束位置的概率$p^{o,s}$和$p^{o,e}$：

[0106] $$p^{o,s} = \mathrm{softmax}(W_{o,2}r_o) \quad (9)$$

[0107] $$p^{o,e} = \mathrm{softmax}(W_{o,3}r_o) \quad (10)$$

[0108] 其中，$W_{o,1} \in \mathbb{R}^{d\times d}, W_{o,2} \in \mathbb{R}^{1\times d}, W_{o,3} \in \mathbb{R}^{1\times d}$是可训练的权重；

[0109] 使用公式(11)计算意见词的开始或结束位置的损失$\mathcal{L}_O$：

[0110] $$\mathcal{L}_O = -\sum_{i=1}^{n} y_i^{o,s}\log p_i^{o,s} - \sum_{i=1}^{n} y_i^{o,e}\log p_i^{o,e} \quad (11)$$

[0111] 其中，$y^{o,s}$和$\mathbf{y}^{o,e} \in \mathbb{R}^n$分别是对于句子中首个未掩码的方面词对应的意见词开始与结束位置的真实值。

[0112] 3、关于情感分类模块：

[0113] 为准确识别情感，模型需要更充分地考虑方面词、意见词以及上下文的信息，由此使用多头自注意力机制融合方面词、意见词以及上下文的信息，这一处理过程如下所示：

[0114] 使用公式(12)获取情感表示r_s：

[0115] $$r_s = \mathrm{LN}(h_x + \mathrm{MultiHead}(h_x, r_a, r_o)) \quad (12)$$

[0116] 其中，LN、MultiHead分别代表层归一化(Layer Normalization)模块、多头注意力(Multihead Attention)网络，其参数中查询(query)、键(key)、值(value)分别为h_x、r_a、r_o；

[0117] 使用公式(13)计算g_s：

[0118] $$g_s = \mathrm{MP}(r_s) \quad (13)$$

[0119] 其中，$\mathbf{g_s} \in \mathbb{R}^{d\times n}$代表，MP代表最大池化(Max Pooling)；

[0120] 使用公式(14)计算p^s：

[0121] $$p^s = \mathrm{softmax}(W_s g_s + b_s) \quad (14)$$

[0122] 其中，p^s代表情感概率，$W_s \in \mathbb{R}^{1\times d}$为可训练的权重，$b_s$为偏置项；

[0123] 使用公式(15)表达情感分类损失：

[0124] $$\mathcal{L}_S = -\sum_{i=1}^{3} y_i^s \log p_i^s \quad (15)$$

[0125] 其中，$\mathbf{y}^s \in \mathbb{R}^3$是真实的情感极性标签。

[0126] 4、关于方面词存在探测模块：

[0127] 这一模块是为了探测经过掩码之后的上下文是否仍存在方面词，如果所有的方面词均被掩码，其标签为False，否则为True。该模块设计过程如下：

[0128] 使用公式(16)获取r_e：

[0129] $$r_e = h_{cls} \oplus MP(r_o) \oplus MP(r_a) \quad (16)$$

[0130] 其中，$r_e \in \mathbb{R}^{3d\times 1}$是中间变量，代表方面词存在探测的表示，⊕表示拼接操作；

[0131] 使用公式(17)计算p^e：

[0132] $$p^e = \mathrm{softmax}(W_e r_e + b_e) \quad (17)$$

[0133] 其中，p^e代表方面词的存在概率，$W_e \in \mathbb{R}^{2\times 3d}$和$b_e$分别是可训练的权重和偏置；

[0134] 使用公式(18)计算二元交叉熵损失$\mathcal{L}_E$：

[0135] $$\mathcal{L}_E = -\sum_{i=1}^{2} y_i^e \log p_i^e \quad (18)$$

[0136] 其中，$y^e \in \mathbb{R}^2$为方面词存在真实标签，p^e为预测的方面词存在概率。

[0137] 5、模型目标函数

[0138] 目标函数是为了最小化以下损失，表示为：

[0139] $$\mathcal{L}_T = \alpha\mathcal{L}_A + \beta\mathcal{L}_O + \gamma\mathcal{L}_S + \delta\mathcal{L}_E \quad (19)$$

[0140] 其中，α，β，γ以及δ分别是调整$\mathcal{L}_A$、$\mathcal{L}_O$、$\mathcal{L}_S$、$\mathcal{L}_E$损失影响因子的超参数。

[0141] 将其他所有方面词掩码掉可以在AAI阶段削弱他们带来的干扰，基于这个想法，本发明设计了一个有效的推理算法如下：

[0142] 算法 1： COM-MRC 推理算法

[0143]

输入：句子 x

输出：三元组 $T=\{(a,o,s)\}_n$

1. 初始化 $T,A=\{\},\{\}$ 以及 $q,c,c_{max}=q_0,0,c_0$
2. $F(q,x)\rightarrow e,a$ // 根据查询 q以及上下文x得到存在标示位e以及方面词a
3. while $e==True \&\& c<c_{max}$ do
4. $A\leftarrow A\cup\{a\}$ // 将方面词 a 并入集合A中
5. $A\leftarrow$ Merge (A) // 合并A中相邻或相交区间
6. $F(q,Mask(x,A))\rightarrow e,a$ // 根据查询 q以及将A中的方面词掩码得到的上下文
7. // Mask(x,A)得到存在标示位e以及方面词a
8. $c\leftarrow c+1$ // 避免无限循环
9. end while
10. for $a_i\in A$ do
11. // 根据查询 q以及掩码其他所有方面词的上下文得到意见词O集合以及情感s
12. $F(q,Mask(x,A-\{a_i\}))\rightarrow O,s$
13. for $o_j\in O$ do
14. $T\leftarrow T\cup\{(a_i,o_j,s)\}$ // 将三元组并入三元组集合T中
15. end for
16. end for
17. return T

[0144] 在该推理算法中，第2行至第9行为方面词推理阶段，第10行至第16行为方面词附属推理阶段。q_0是一个固定的查询，如公式3所示，c_0是一个固定的常数以阻止可能的无限循环，在这里我们设定为10，因为一句话通常不超过10个方面词。F代表模型函数，其输入包括一个查询和一个掩码上下文。Mask(x,A)代表将属于A中所有的方面词掩码掉的上下文x。Merge(A)代表将集合A中所有的相邻或相交区间进行合并。

[0145] 如图3的推理部分所示，对于句子“Nice ambience,but highly overrated place”，我们在方面词推理阶段识别出“ambience”、“place”两个方面词，然后在意见词推理阶段我们识别出“ambience”的意见词为“Nice”，情感为积极的；识别出“place”的意见词为“overrated”，情感为消极的。图1展示了一个ASTE任务示例，并且展示了传统MRC与本发明方法的主要不同之处。可见，本发明的方法有效的解决了以往MRC方法面临的方面词干扰问题。

[0146] 以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换，但这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

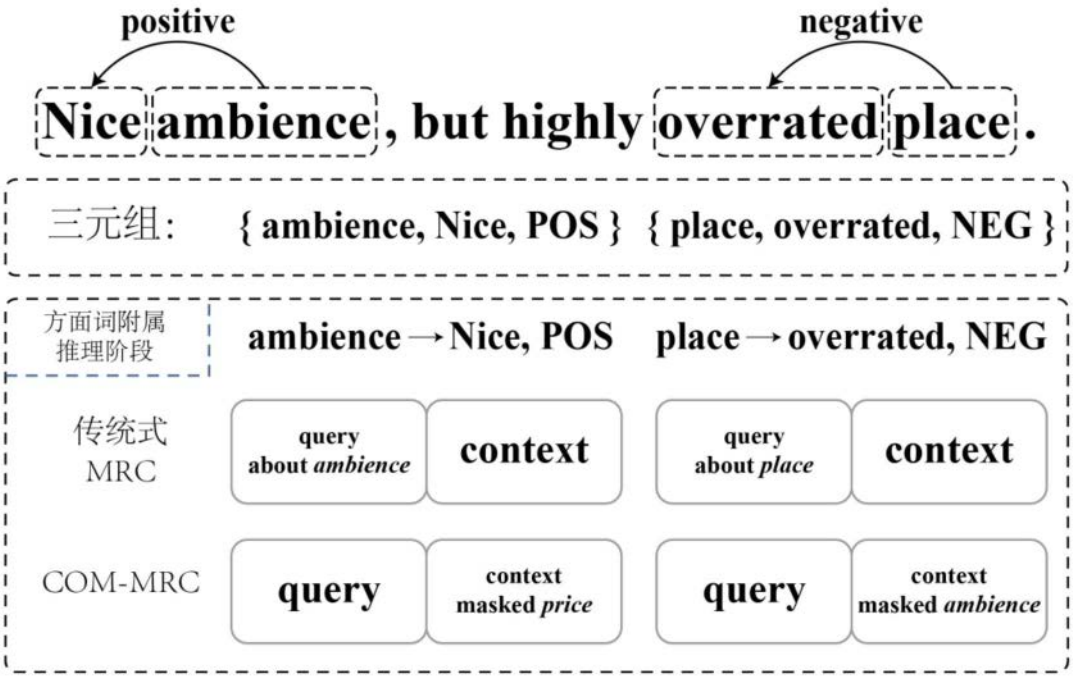

positive
negative
Nice ambience , but highly overrated place .
三元组: { ambience, Nice, POS } { place, overrated, NEG }
方面词附属推理阶段
ambience → Nice, POS
place → overrated, NEG
传统式MRC
query about ambience
context
query about place
context
COM-MRC
query
context masked price
query
context masked ambience

图1

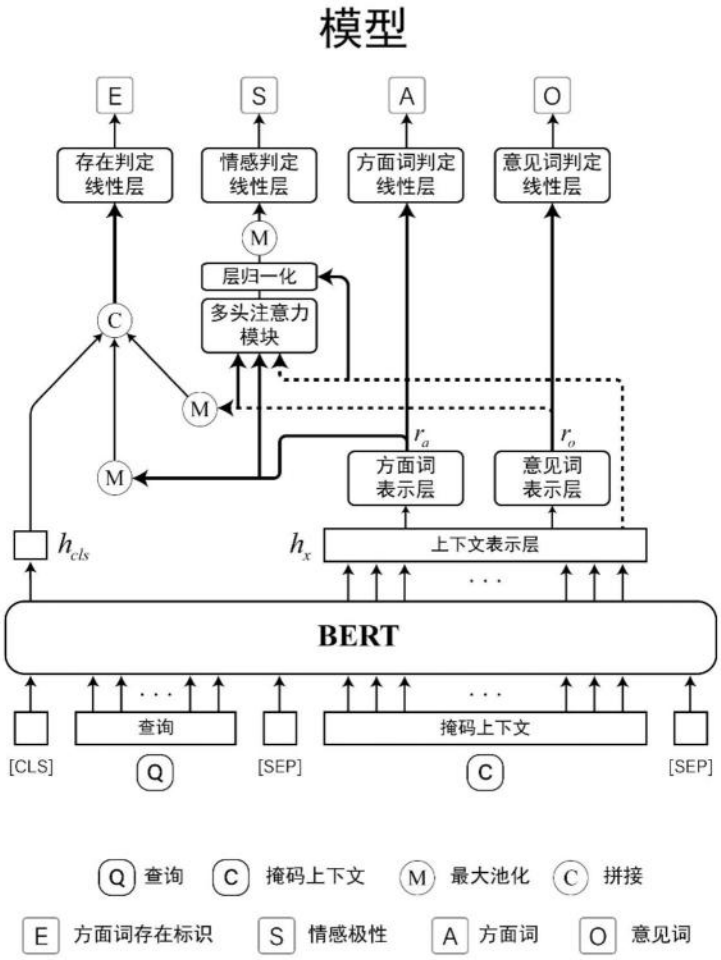

模型
E
S
A
O
存在判定线性层
情感判定线性层
方面词判定线性层
意见词判定线性层
M
层归一化
多头注意力模块
C
M
M
r_a
r_o
方面词表示层
意见词表示层
h_{cls}
h_x
上下文表示层
BERT
查询
掩码上下文
[CLS]
Q
[SEP]
C
[SEP]
Q 查询
C 掩码上下文
M 最大池化
C 拼接
E 方面词存在标识
S 情感极性
A 方面词
O 意见词

图2

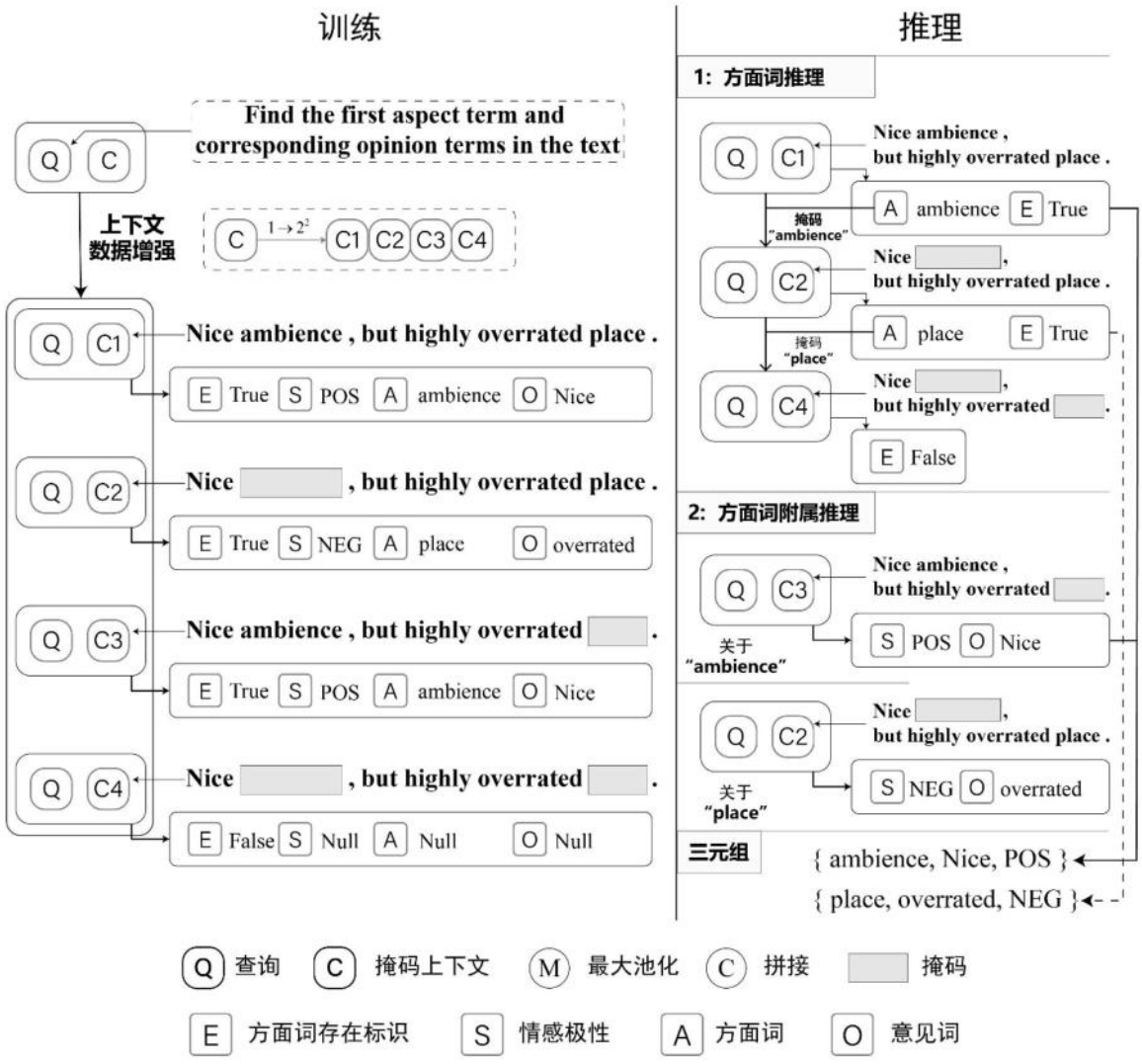

训练
推理
Find the first aspect term and corresponding opinion terms in the text
上下文数据增强
Nice ambience , but highly overrated place .
Nice , but highly overrated place .
Nice ambience , but highly overrated .
Nice , but highly overrated .
1: 方面词推理
2: 方面词附属推理
三元组
{ ambience, Nice, POS }
{ place, overrated, NEG }
掩码 "ambience"
掩码 "place"
关于 "ambience"
关于 "place"
Q 查询
C 掩码上下文
M 最大池化
C 拼接
掩码
E 方面词存在标识
S 情感极性
A 方面词
O 意见词

图3

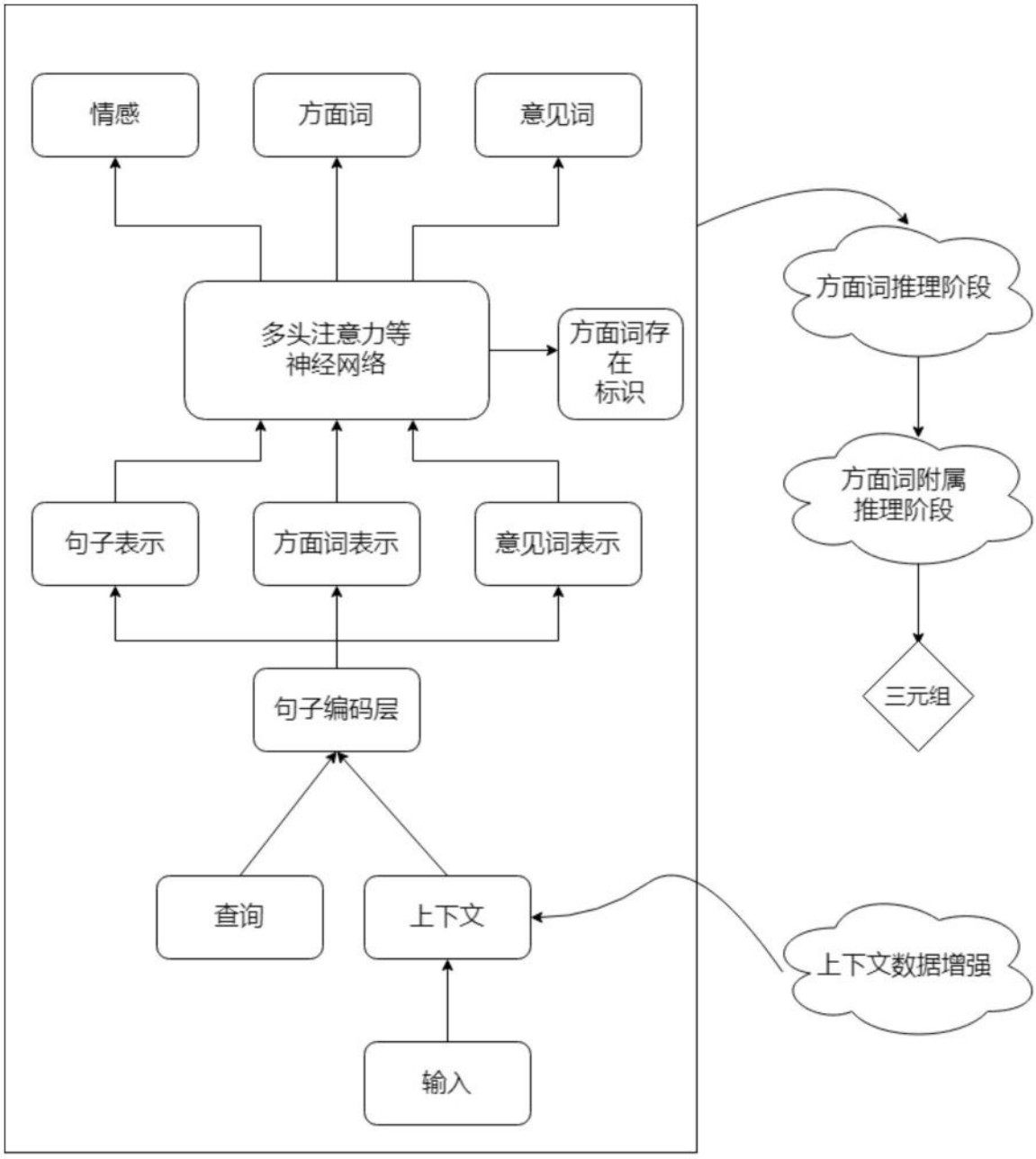

情感
方面词
意见词
多头注意力等
神经网络
方面词存
在
标识
句子表示
方面词表示
意见词表示
句子编码层
查询
上下文
输入
方面词推理阶段
方面词附属
推理阶段
三元组
上下文数据增强

图4