(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116522165 B
(45)授权公告日 2024.04.02

(21)申请号 202310761055.3
(22)申请日 2023.06.27
(65)同一申请的已公布的文献号
申请公布号 CN 116522165 A
(43)申请公布日 2023.08.01
(73)专利权人 武汉爱科软件技术股份有限公司
地址 430000 湖北省武汉市东湖新技术开发区民院路38号龙安·港汇城A单元11层02室
(72)发明人 陈宏伟 涂麟曦
(74)专利代理机构 武汉智汇为专利代理事务所(普通合伙) 42235
专利代理师 李恭渝
(51)Int.Cl.
G06F 40/295(2020.01)
G06F 18/22(2023.01)
G06F 40/30(2020.01)
G06F 18/2415(2023.01)

(56)对比文件
CN 111259127 A,2020.06.09
CN 111723575 A,2020.09.29
CN 113673225 A,2021.11.19
CN 114329225 A,2022.04.12
CN 114386421 A,2022.04.22
CN 114579731 A,2022.06.03
CN 114896397 A,2022.08.12
CN 115292447 A,2022.11.04
CN 115374778 A,2022.11.22
CN 115408494 A,2022.11.29
CN 115470871 A,2022.12.13
CN 115630632 A,2023.01.20
CN 115687939 A,2023.02.03
CN 115712713 A,2023.02.24
CN 115759104 A,2023.03.07
CN 116304745 A,2023.06.23
US 2016042061 A1,2016.02.11 (续)

审查员 杨静

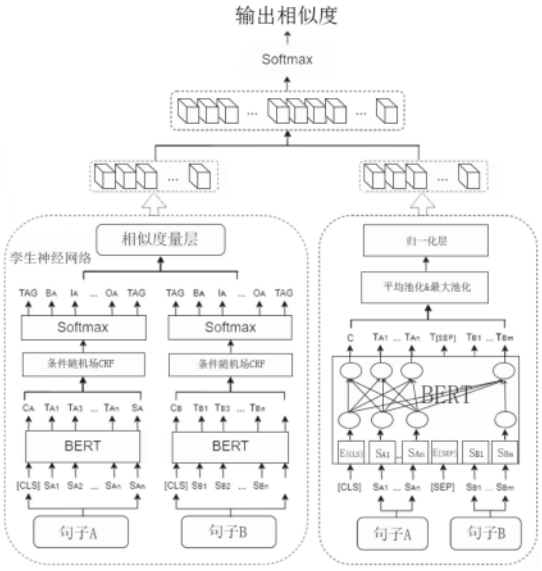
权利要求书2页 说明书10页 附图2页

(54)发明名称

一种基于孪生结构的舆情文本匹配系统及方法

(57)摘要

本发明的一种基于孪生结构的舆情文本匹配系统,包括孪生神经网络模块:用于构造孪生神经网络的编码层,获取命名实体间的第一相似度表征向量;语义交互模块:用于获取第二相似度表征向量;融合模块:用于将第一相似度表征向量和第二相似度表征向量拼接,得到句子对的最终相似度表征向量;匹配模块:用于将最终相似度表征向量通过SoftMax分类函数得到文本匹配结果。本发明通过提取舆情文本的命名实体相似度特征和文本语义相似度特征,将两类特征融合后进行语义相似度计算并分析两舆情文本是否相似,提高舆情文本匹配的准确性和鲁棒性,因为不再是单纯对文本的主题和含义进行匹配,同时考虑了针对同一人物、事物或现象的表述进行匹配。



CN 116522165 B

[转续页]

[接上页]

(56)对比文件

US 2022198146 A1,2022.06.23

陈剑;何涛;闻英友;马林涛.基于BERT模型的司法文书实体识别方法.东北大学学报(自然科学版).2020,(第10期),第16-21页.

谢腾;杨俊安;刘辉.基于BERT-BiLSTM-CRF模型的中文实体识别.计算机系统应用.2020,(第07期),第52-59页.

向军毅 等.基于BERTCA的新闻实体与正文语义相关度计算模型.《第十九届中国计算语言学大会论文集》.2020,第288-300页.

Leonidas Tsekouras 等.A Graph-based Text Similarity Measure That Employs Named Entity Information .《Proceedings of Recent Advances in Natural Language Processing》.2017,第765-771页.

1.一种基于孪生结构的舆情文本匹配的系统,其特征在于包括

孪生神经网络模块:用于构造孪生神经网络的编码层,从而提取出句子对中的命名实体信息,并对提取出的命名实体进行相似度计算,获取命名实体间的第一相似度表征向量;

语义交互模块:用于获取句子对在语义方面的第二相似度表征向量, 第二相似度向量表示文本语义相似度特征;

所述语义交互模块,包括交互模块的编码层、交互模块的池化层和交互模块的归一化层,所述交互模块的编码层,将句子对送入BERT前,需要在句子的头部加入[CLS]标识符,并在两句之间插入[SEP]标识符进行切分,将拼接好后的句子T＝{$[CLS],S_{A1},S_{A2},\ldots,S_{An}$,[SEP],$S_{B1},S_{B2},\ldots,S_{Bm}$}送入BERT模型进行微调,通过BERT层的编码为句子中每个位置上的字符引入上下文信息从而获取词性状态以进行表征向量,输出E_T＝{$T_{[CLS]},T_{A1},T_{A2},\ldots,T_{An},T_{[SEP]},T_{B1},T_{B2},\ldots,T_{Bm}$},即句子对的向量化表达;所述交互模块的池化层,通过BERT得到的句向量E_T通过池化层来提取重要特征缩小维度;所述交互模块的归一化层,句向量E_T经过层归一化后的输出结果为交互模块获取到的句子对的第二相似度表征向量;

所述孪生神经网络模块,具体利用BERT+CRF方法构造孪生神经网络的编码层,孪生神经网络模块包括两个相同的神经网络建立的耦合三层架构,孪生神经网络的每个神经网络包括输入层、特征提取层,两个相同的神经网络共享相似度度量层,其中每个神经网络的输入层输入需进行匹配的句子对的一个句子,特征提取层将输入的句子对的句子嵌入至高维度空间得到句子对的的句子的向量表征,相似度度量层通过数学公式对提取出的句子对的句子的向量表征进行相似度计算,得到句子对的第一相似度表征向量;将句子对的A、B句子送入孪生神经网络前,需要在句子的头部加入[CLS]标识符,得到A、B句子对的A句子向量$T_A=\{[CLS],S_{A1},S_{A2},\ldots,S_{An}\}$和B句子向量$T_B=\{[CLS],S_{B1},S_{B2},\ldots,S_{Bm}\}$;将$T_A$和$T_B$送入BERT进行微调,通过BERT层的编码为句子中每个位置上的字符引入上下文信息从而获取词性状态以进行表征向量,所有BERT的输出将作为CRF层的输入;

包括标注单元,孪生神经网络模块的训练集即句子对采用BIO方法对实体进行标注,B表示字符处于一个实体的开始,I表示字符处于该实体的内部位置,O表示实体外部的不被关注的非实体字符;舆情文本需重点关注文本中的人名PER、地名GEO、以及组织ORG,故训练集的实体标签有B-PER,I-PER,B-GEO,I-GEO,B-ORG,I-ORG,O这7种类型的标签;B为begin的简写,I为inside的简写,O为outside的简写;

融合模块:用于将第一相似度表征向量和第二相似度表征向量拼接,得到句子对的最终相似度表征向量;

匹配模块:用于将最终相似度表征向量通过SoftMax分类函数得到文本匹配结果。

2.基于孪生结构的舆情文本匹配的方法,应用如权利要求1所述的基于孪生结构的舆情文本匹配的系统,其特征在于,包括如下步骤:

构造孪生神经网络的编码层,从而提取出句子对中的命名实体信息,并对提取出的命名实体进行相似度计算,获取命名实体间的第一相似度表征向量;

通过语义交互模块获取句子对在语义方面的第二相似度表征向量,第二相似度向量表示文本语义相似度特征;

将第一相似度表征向量和第二相似度表征向量拼接,得到句子对的最终相似度表征向量;

将最终相似度表征向量通过SoftMax分类函数得到文本匹配结果。

3.一种计算机可读的存储介质,其特征在于,所述计算机可读的存储介质包括存储的程序,其中,所述程序运行时执行上述权利要求2中所述的基于孪生结构的舆情文本匹配的方法。

4.一种电子装置,包括存储器和处理器,其特征在于,所述存储器中存储有计算机程序,所述处理器被设置为通过所述计算机程序执行所述权利要求2中所述的基于孪生结构的舆情文本匹配的方法。

一种基于孪生结构的舆情文本匹配系统及方法

技术领域

[0001] 本发明属于自然语言处理技术领域,具体涉及一种基于孪生结构的舆情文本匹配系统及方法。

背景技术

[0002] 目前舆情本文匹配方法的核心问题是解决文本数据相似度判断的问题,只有当文本数据相似度判断准确了,舆情文本系统的匹配准确率才能提升。在以往传统方法中,需要大量人力和时间进行人为判断、标注和去除相似的舆情文本。因此需要一种智能化的舆情文本匹配系统,提炼重要信息,提高文本分析的效率。舆情文本匹配在舆情分析、舆情预警中发挥着至关重要地作用,舆情文本匹配的准确率关乎着后续舆情研判的准确与否。

[0003] 目前对于舆情文本匹配的计算大多采用两种方式,一种是基于传统的文本匹配算法,另一种是基于深度学习的文本匹配算法。传统的文本匹配算法一般可分为基于字符串的方法,基于统计的方法和基于知识库的方法。传统的文本匹配算法大多都只能计算出文本表层的含义,难以挖掘出文本深层含义。随着自然语言处理任务的需求越来越广泛,基于传统的方法始终无法突破语义相似度计算任务的瓶颈,故逐渐被基于深度学习的语义相似度算法取代。基于深度学习的文本匹配算法可以理解到文本的深层含义,使模型效果更好,但由于研究时间不长,模型的准确性仍待提升。在2013年提出的生成分布式词向量方法,即word2vec,该方法根据一定范围内的上下文预测出来文本中每个单词的词向量,然后生成的词向量被拼接后,能够表示一定的语义信息;但每个词所依赖的上下文范围是有限的,因此每个词向量表达句语义信息也是局部有限的。2014年又提出了doc2vec方法,该方法用于文档文本的向量化表示,文档与单词不同之处在于,文档没有像单词与单词之间的逻辑结构,其是一个整体的文本数据。以上两种方法所生成的向量均为静态的,即无法根据文本语境的不同而动态变化,从而影响了方法的准确率和性能。

[0004] 近几年BERT方法的提出,给自然语言处理领域带来了很大影响,BERT方法结合了自注意力机制,并提出了掩盖语言模型任务和下文预测任务两种十分新颖且有效的预训练目标,为方法的性能带来极大的提升,成为目前最常用的生成动态词向量的方法之一。舆情文本匹配比起一般的文本匹配具有更高的难度,它不仅仅需要判断两文本在语义上是否相似,还需判断两文本是否是针对同一人物、事物或现象所表达的信念、态度、意见和情绪等等。现有的文本匹配算法一般只考虑文本字符的匹配或文本含义匹配,即当两文本有许多相似字符或两文本表达相同主题或相同含义时则判断为相似,未具体到人物或事件层面,故本发明提出一种基于孪生结构的舆情文本匹配方法以使舆情场景的文本匹配在准确率和鲁棒性方面得到进一步提升。

发明内容

[0005] 针对舆情文本匹配比起一般的文本匹配具有更高的难度,不仅仅需要判断两文本在语义上是否相似,还需判断两文本是否是针对同一人物、事物或现象所表达的信念、态

度、意见和情绪等,因此,基于舆情场景的文本匹配方法在准确率和鲁棒性方均要求更高,不仅要判断两文本在语义上是否相似,还需要判断两文本是否为针对同一人物、事物或现象的表述。

[0006] 为了克服上述现有技术的不足,本发明旨在提供一种基于孪生结构的舆情文本匹配系统及方法。

[0007] 根据本发明的第一方面,提供一种基于孪生结构的舆情文本匹配的系统,包括

[0008] 孪生神经网络模块:用于构造孪生神经网络的编码层,提取出句子对中的命名实体信息,并对提取出的命名实体进行相似度计算,获取命名实体间的第一相似度表征向量;

[0009] 语义交互模块:用于获取句子对在语义方面的第二相似度表征向量;

[0010] 融合模块:用于将第一相似度表征向量和第二相似度表征向量拼接,得到句子对的最终相似度表征向量;

[0011] 匹配模块:用于将最终相似度表征向量通过SoftMax分类函数得到文本匹配结果。

[0012] 在本发明的一种示例性实施例中,所述孪生神经网络模块,具体利用BERT+CRF方法构造孪生神经网络的编码层,包括两个相同或相似的神经网络建立的耦合三层架构,分别是输入层、特征提取层和相似度度量层,其中输入层输入需进行匹配的句子对,特征提取层将输入的句子对样本嵌入至高纬度空间得到句子对两个样本的表征向量,相似度度量层通过数学公式对提取出的两个样本的表征向量进行相似度计算,得到句子对的第一相似度表征向量。

[0013] 在本发明的一种示例性实施例中,所述孪生神经网络模块的BERT模型

[0014] 还包括掩码语言模型任务单元,(采用BERT层的掩码语言模型任务获取输入句子对语句中词级别的文本特征),在训练的输入层中随即掩盖部分字符,然后利用剩余未被掩盖的字符来预测这些掩盖的字符,通过该方式的训练,可使模型充分学习到输入语句中词级别的文本特征,再将BERT层输出的特征向量输入至CRF层;

[0015] 还包括下文预测任务单元,用于判断输入的句子对的A句子和B句子是否上下问相关,从而使模型学习到两个文本之间的关系,解决句子层面的问题;再将BERT层输出的特征向量输入至CRF层;

[0016] 在本发明的一种示例性实施例中,所述孪生神经网络模块的CRF模型还包括数据集中标签之间的转移概率单元,CRF层通过学习数据集中标签之间的转移概率,从而修正BERT层的输出,从而保证预测标签的合理性;

[0017] 还包括标注单元,由于需提取出句子对中的命名实体,训练集即句子对采用BIO方法对实体进行标注,B(begin)表示该字符处于一个实体的开始,I(inside)表示该字符处于该实体的内部位置,O(outside)表示实体外部的不被关注的非实体字符;对于舆情文本需重点关注文本中的人名(PER)、地名(GEO)、以及组织(ORG),故训练集的实体标签有B-PER,I-PER,B-GEO,I-GEO,B-ORG,I-ORG,O这7种类型的标签;

[0018] 还包括获取词性状态以进行表征向量单元,将句子对送入孪生神经网络前,需要在句子的头部加入[CLS]标识符,得到A、B句子对的A句子$T_A = \{[CLS], S_{A1}, S_{A2}, \ldots, S_{An}\}$和$T_B = \{[CLS], S_{B1}, S_{B2}, \ldots, S_{Bm}\}$;将$T_A$和$T_B$送入BERT进行微调,通过BERT层的编码为句子中每个位置上的字符引入上下文信息从而获取词性状态以进行表征向量,所有BERT的输出将作为CRF层的输入;

[0019] 在本发明的一种示例性实施例中,所述语义交互模块,具体基于BERT采用下文预测任务以学习文本间的句子关系特征,包括交互模块的编码层、交互模块的池化层和交互模块的归一化层,所述交互模块的编码层,将句子对送入BERT前,需要在句子的头部加入[CLS]标识符,并在两句之间插入[SEP]标识符进行切分。将拼接好后的句子$T=\{[CLS],S_{A1},S_{A2},\ldots,S_{An},[SEP],S_{B1},S_{B2},\ldots,S_{Bm}\}$送入BERT模型进行微调,输出$E_T=\{T_{[CLS]},T_{A1},T_{A2},\ldots,T_{An},T_{[SEP]},T_{B1},T_{B2},\ldots,T_{Bm}\}$,即句子对的向量化表达;

[0020] 所述交互模块的池化层,通过BERT得到的句向量E_T通过池化层来提取重要特征缩小维度;

[0021] 所述交互模块的归一化层,句向量E_T经过层归一化后的输出结果为交互模块获取到的句子对的第二相似度表征向量。

[0022] 在本发明的一种示例性实施例中,所述匹配模块中,具体SoftMax分类函数如下,$p(y=j)$代表的含义为样本向量 x 属于第 j 个分类的概率,其中W为权重系数,k表示有k个类别:

[0023] $$p(y=j)=\frac{e^{x^T W_j}}{\sum_{k=1}^{K} e^{x^T W_k}}$$

[0024] 将最终相似度表征向量E_{All}输入至softmax函数中,$E_{All}=Concat(E_{SNN},E'_T)$,其中$E_{SNN}$为所述孪生神经网络模块的输出,$E'_T$为所述交互模块的输出,$E_{All}$为上述 softmax 函数中的 x;得到的最终结果p在[0,1]区间中,假设设置文本相似的阈值为0.5,则当$p\geq 0.5$时,则认为两文本匹配,否则两文本不匹配。

[0025] 根据本发明的第二方面,提供一种基于孪生结构的舆情文本匹配的方法,应用所述的基于孪生结构的舆情文本匹配的系统,包括如下步骤:

[0026] 构造孪生神经网络的编码层,从而提取出句子对中的命名实体信息,并对提取出的命名实体进行相似度计算,获取命名实体间的第一相似度表征向量;

[0027] 获取句子对在语义方面的第二相似度表征向量;

[0028] 将第一相似度表征向量和第二相似度表征向量拼接,得到句子对的最终相似度表征向量;

[0029] 将最终相似度表征向量通过SoftMax分类函数得到文本匹配结果。

[0030] 在本发明的一种示例性实施例中,所述构造孪生神经网络的编码层,具体利用BERT+CRF方法构造孪生神经网络的编码层,包括两个相同或相似的神经网络建立的耦合三层架构,分别是输入层、特征提取层和相似度度量层,其中输入层输入需进行匹配的句子对,特征提取层将输入的句子对样本嵌入至高纬度空间得到句子对两个样本的表征向量,相似度度量层通过数学公式对提取出的两个样本的表征向量进行相似度计算,得到句子对的第一相似度表征向量。

[0031] 在本发明的一种示例性实施例中,所述构造孪生神经网络的编码层,从而提取出句子对中的命名实体信息,并对提取出的命名实体进行相似度计算,获取命名实体间的第一相似度表征向量,具体还包括:

[0032] 采用BERT层的掩码语言模型任务获取输入句子对语句中词级别的文本特征,再将BERT层输出的特征向量输入至CRF层;

[0033] CRF层通过学习数据集中标签之间的转移概率,从而修正BERT层的输出;

[0034] 训练集即句子对采用BIO方法对实体进行标注,B(begin)表示该字符处于一个实体的开始,I(inside)表示该字符处于该实体的内部位置,O(outside)表示实体外部的不被关注的非实体字符;对于舆情文本需重点关注文本中的人名(PER)、地名(GEO)、以及组织(ORG),故训练集的实体标签有B-PER,I-PER,B-GEO,I-GEO,B-ORG,I-ORG,O这7种类型的标签;

[0035] 将句子对送入孪生神经网络前,需要在句子的头部加入[CLS]标识符,得到A、B句子对的A句子向量$T_A = \{[CLS], S_{A1}, S_{A2}, \ldots, S_{An}\}$和$T_B = \{[CLS], S_{B1}, S_{B2}, \ldots, S_{Bm}\}$;将$T_A$和$T_B$送入BERT进行微调,通过BERT层的编码为句子中每个位置上的字符引入上下文信息从而获取词性状态以进行表征向量,所有BERT的输出将作为CRF层的输入。

[0036] 在本发明的一种示例性实施例中,所述获取句子对在语义方面的第二相似度表征向量,具体包括:具体基于BERT采用下文预测任务以学习文本间的句子关系特征;将句子对送入BERT前,需要在句子的头部加入[CLS]标识符,并在两句之间插入[SEP]标识符进行切分。将拼接好后的句子$T = \{[CLS], S_{A1}, S_{A2}, \ldots, S_{An}, [SEP], S_{B1}, S_{B2}, \ldots, S_{Bm}\}$送入BERT模型进行微调,输出$E_T = \{T_{[CLS]}, T_{A1}, T_{A2}, \ldots, T_{An}, T_{[SEP]}, T_{B1}, T_{B2}, \ldots, T_{Bm}\}$,即句子对的向量化表达;通过BERT得到的句向量$E_T$通过池化层来提取重要特征缩小维度;句向量$E_T$经过层归一化后的输出结果为交互模块获取到的句子对的第二相似度表征向量。

[0037] 根据本发明的第三方面,提供一种计算机可读的存储介质,所述计算机可读的存储介质包括存储的程序,其中,所述程序运行时执行上述的基于孪生结构的舆情文本匹配的方法。

[0038] 根据本发明的第四方面,提供一种电子装置,包括存储器和处理器,所述存储器中存储有计算机程序,所述处理器被设置为通过所述计算机程序执行所述的基于孪生结构的舆情文本匹配的方法。

[0039] 本发明所构思的以上技术方案与现有技术相比,能够取得下列有益效果:

[0040] 本发明的基于孪生结构的舆情文本匹配系统和方法,该系统分为两个主要模块,分别为基于BERT+CRF的孪生神经网络模块和基于BERT的语义交互模块。孪生神经网络模块利用BERT+CRF方法构造孪生神经网络的编码层,从而提取出句子对中的命名实体信息包括人名,地名等,并对提取出的命名实体进行相似度计算,获取命名实体间的相似度特征(表征向量)。基于BERT的语义交互模块可获取句子对在语义方面的相似度特征(表征向量)。本发明通过以上两个模块提取舆情文本的命名实体相似度特征和文本语义相似度特征,将两类特征融合后进行语义相似度计算并分析两舆情文本是否相似,提高舆情文本匹配的准确性和鲁棒性,因为不再是单纯对文本的主题和含义进行匹配,同时考虑了针对同一人物、事物或现象的表述进行匹配。

附图说明

[0041] 图1为本发明基于孪生结构的舆情文本匹配系统结构示意图。

[0042] 图2为本发明孪生神经网络模块的BERT模型的输入表征向量图。

[0043] 图3为本发明孪生神经网络模块的训练集的具体标签形式图。

具体实施方式

[0044] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,均属于本发明保护的范围。

[0045] 实施例一

[0046] 结合图1所示,本实施例提供一种提供一种基于孪生结构的舆情文本匹配的系统,包括:孪生神经网络模块,用于构造孪生神经网络的编码层,提取出句子对中的命名实体信息,并对提取出的命名实体进行相似度计算,获取命名实体间的第一相似度表征向量;语义交互模块,用于获取句子对在语义方面的第二相似度表征向量;融合模块,用于将第一相似度表征向量和第二相似度表征向量拼接,得到句子对的最终相似度表征向量;匹配模块,用于将最终相似度表征向量通过SoftMax分类函数得到文本匹配结果。

[0047] 在一种示例性实施例中,所述孪生神经网络模块,具体利用BERT+CRF方法(即BERT模型+CRF模型)构造孪生神经网络的编码层,包括两个相同或相似的神经网络(具体BERT模型+CRF模型)建立的耦合三层架构,该耦合三层架构的天然优势使其非常适用于解决相似度匹配问题。三层架构分别是输入层、特征提取层和相似度度量层,其中输入层输入需进行匹配的句子对样本,特征提取层将输入的句子对样本嵌入至高纬度空间得到两个句子对样本的表征向量,相似度度量层通过数学公式对提取出的两个样本的表征向量进行相似度计算,得到句子对的第一相似度表征向量,一般可以采用欧式距离、余弦距离或杰卡德距离等方法计算两样本的相似度。

[0048] 具体地,BERT模型采用多层Transfomer编码器作为其网络层,从而能够深度挖掘文本中的重要特征,捕捉更长距离的上下文信息。BERT是一个多任务模型,预训练好的BERT模型能够完成各式各样的下游任务。该模型的输入既可以为单个语句,也可以是文本。文本输入时,需要将文本序列的首部添加一个特殊分类符号[CLS],然后在每句话的结束位置添加一个特殊符号[SEP]作为句子的分隔符和结束符。文本中的每个字符首先通过word2vec模型进行向量初始化形成原始表征向量。为了区分字符来源,需要添加一个片段归属信息嵌入来区分该字符是来自于句子对的句子A还是句子B。最后,为了(是)模型学会句子中各个字符的位置信息对句子含义的影响,还需要嵌入一个位置向量。故最终BERT模型的输入表征向量由字嵌入、片段归属信息嵌入和位置嵌入三部分相加而成,如图2所示。

[0049] BERT模型的预训练任务由两个无监督学习子任务组成,分别是掩码语言模型和下文预测任务。掩码语言模型是指在训练的输入层中随即掩盖部分字符,然后利用剩余未被掩盖的字符来预测这些掩盖的字符,通过该方式的训练,可使模型充分学习到输入语句中词级别的文本特征。下文预测任务是让模型判断输入的两句子是否上下问相关,从而使模型学习到两个文本之间的关系,解决句子层面的问题。每个字符的通过大量的无监督语料进行上述两种任务的充分训练后,学习到文本的语言特征并输出具有更深层次表达的字符向量编码。在下游任务中,可直接利用训练好的模型参数对文本进行向量化。

[0050] 在一种示例性实施例中,所述孪生神经网络模块的BERT模型,还包括掩码语言模型任务单元,(采用BERT层的掩码语言模型任务获取输入句子对语句中词级别的文本特征),在训练的输入层中随即掩盖部分字符,然后利用剩余未被掩盖的字符来预测这些掩盖

的字符,通过该方式的训练,可使模型充分学习到输入语句中词级别的文本特征,再将BERT层输出的特征向量输入至CRF层;还包括下文预测任务单元,用于判断输入的句子对的A句子和B句子是否上下问相关,从而使模型学习到两个文本之间的关系,解决句子层面的问题;再将BERT层输出的特征向量输入至CRF层。

[0051] 在本发明的一种示例性实施例中,所述孪生神经网络模块的CRF模型,还包括数据集中标签之间的转移概率单元,CRF层通过学习数据集中标签之间的转移概率,从而修正BERT层的输出,从而保证预测标签的合理性,修正bert层的输出,比如之前BERT输出的是向量X,修正后输出为X´;还包括标注单元,由于需提取出句子对中的命名实体训练集即句子对采用BIO方法对实体进行标注,B(begin)表示该字符处于一个实体的开始,I(inside)表示该字符处于该实体的内部位置,O(outside)表示实体外部的不被关注的非实体字符;对于舆情文本需重点关注文本中的人名(PER)、地名(GEO)、以及组织(ORG),故训练集的实体标签有B-PER,I-PER,B-GEO,I-GEO,B-ORG,I-ORG,O这7种类型的标签;还包括获取词性状态以进行表征向量单元,将句子对的A、B句子送入孪生神经网络前,需要在句子的头部加入[CLS]标识符,得到A、B句子对的A句子向量$T_A = \{[CLS], S_{A1}, S_{A2}, \ldots, S_{An}\}$和B句子向量$T_B = \{[CLS], S_{B1}, S_{B2}, \ldots, S_{Bm}\}$;将$T_A$和$T_B$送入BERT进行微调,通过BERT层的编码为句子中每个位置上的字符引入上下文信息从而获取词性状态以进行表征向量,所有BERT的输出将作为CRF层的输入;还可以包括预处理单元,对于孪生神经网络模块的训练集即句子对作为模型的输入,对该输入句子对的文本进行清洗和停用词去除,并采用停用词表对整个文本进行过滤(从而降低文本长度,提高模型的计算效率),采用直接截断的方式对输入的文本长度进行限制。

[0052] 总之,孪生神经网络模块利用BERT+CRF方法构造孪生神经网络的编码层,从而提取出句子对中的命名实体信息,并对提取出的命名实体进行相似度计算。孪生神经网络模块首先采用BERT层的掩码语言模型任务来获取输入语句中词级别的文本特征。再将BERT层输出的特征向量输入至CRF层,CRF层可以通过学习数据集中标签之间的转移概率从而修正BERT层的输出,从而保证预测标签的合理性。具体如下:

[0053] 舆情文本句子对句子A和句子B,即句子A和句子B为需要判断是否相似的句子对,即孪生神经网络模块的训练集,则该句子对作为模型的输入,需首先对该输入文本进行清洗和停用词去除。文本清洗即对文本中的冗余信息和错误信息进行处理,将空白符号或表情符号等不重要的信息删除,将文本中的繁体字转为简体字,其次将文本中的字符格式统一为半角格式方便后续的文本表征向量。对于文本中的语气词或一些不重要的词可直接删除,并采用停用词表对整个文本进行过滤,从而降低文本长度,提高模型的计算效率。采用直接截断的方式对输入的文本长度进行限制。处理后的句子A长度为n,句子B长度为m, 则表示为A={WA1,WA2,...,WAn},B={WB1,WB2,...,WBn},其中WAi和WBi分别表示句子A和句子B的第i个字。由于需提取出句子对中的命名实体,训练集采用BIO方法对实体进行标注,B(begin)表示该字符处于一个实体的开始,I(inside)表示该字符处于该实体的内部位置,O(outside)表示实体外部的不被关注的非实体字符。对于舆情文本需重点关注文本中的人名(PER)、地名(GEO)、以及组织(ORG),故训练集的实体标签有B-PER,I-PER,B-GEO,I-GEO,B-ORG,I-ORG,O这7种类型的标签。具体标签形式如图3所示。将句子对送入孪生神经网络前,需要在句子的头部加入[CLS]标识符,得到 $T_A = \{[CLS], S_{A1}, S_{A2}, \ldots, S_{An}\}$和$T_B = \{[CLS], S_{B1}, S_{B2}, \ldots, S_{Bm}\}$。将

T_A和T_B送入BERT进行微调,通过BERT层的编码可以为句子中每个位置上的字符引入上下文信息从而获取词性状态以进行表征向量$E_A = \{T_{[CLS]}, E_{A1}, E_{A2}, \ldots, E_{An}\}$和$E_B = \{T_{[CLS]}, E_{B1}, E_{B2}, \ldots, E_B\}$。$E_{Ai}$表示句子A对应的第i个字的编码向量,$E_{Bi}$文表示句子B对应的第i个字的编码向量。所有BERT的输出将作为CRF层的输入。CRF有两类特征函数,一类是针对观测序列与状态的对应关系(如“我”一般是名词),一类是针对状态间关系(如“动词”后一般跟“名词”)。在BERT+CRF模型中,前一类特征函数的输出由BERT的输出替代,后一类特征函数的输出则为标签转移矩阵T,标签转移矩阵T表示标签之间的转移得分。具体的,BERT层输出的表征向量E_A为一个矩阵,得到每个字符S_{Ai}对应的标签得分分布为E_{Ai},将该矩阵称为发射矩阵。对于句子A,其对应的标签$y_A = (y_{A1}, y_{A2}, \ldots y_{An})$是一条链。句子A的长度为n,共有7种类型的标签,故共有n^7种可能的标记结果,即有n^7种可能的y_A。对于舆情文本需重点关注文本中的人名(PER)、地名(GEO)、以及组织(ORG),故训练集的实体标签有B-PER,I-PER,B-GEO,I-GEO,B-ORG,I-ORG,O这7种类型的标签。标签少或者多取决于具体的应用场景,这里只是表述一般情况。对于字符S_{Ai},其标签得分分布E_{Ai}为一个7维的向量,标签y_{Ai}的得分为$E_{Ai}[y_{Ai}]$,其中y_{Ai}为整数类型,表示标签索引。将所有的$E_{Ai}[y_{Ai}]$加起来得到各个字符节点的分数。根据标签矩阵T,求得y_{Ai-1}到y_{Ai}的转移分数$T[i-1,i]$。最后将所有分数求和得到句子A每个可能的标注结果的得分为$score(y_A) = \sum_{i=1}^{n}(E_{Ai}) + \sum_{i=2}^{n}(T[i-1,i])$。然后利用Softmax函数进行归一化求出每种标注结果的概率$p_A(y|x) = \frac{e^{score(y_A)}}{Z_A}$,其中$Z_A = \sum_{y_A} e^{score(y_A)}$。同理地,句子B每种标注结果的概率为$p_B(y|x) = \frac{e^{score(y_B)}}{Z_B}$,其中$Z_B = \sum_{y_B} e^{score(y_B)}$。

[0054] 取概率最大的标注结果作为该字符的实体标签,并将标签为B的字符作为实体的开头,后面跟着标签I的所有字符拼接在一起组成一个词语作为实体词。将实体词所在字符位置对应的BERT层输出的字符表征向量提取得到E'_A和E'_B,利用余弦算法构造相似度度量层,两向量之间的距离特征计算如下:

[0055] $$E_{SNN} = \cos(E'_A, E'_B) = \frac{E'_A \cdot E'_B}{\left|E'_A\right|\left|E'_B\right|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

[0056] E_{SNN}表示利用孪生神经网络(SEN)获得的句子对的相似度特征矩阵即第一相似度表征向量,该相似度特征,将会进一步与BERT获取到的句子对的交互特征进行再次融合。

[0057] 在一种示例性实施例中,所述语义交互模块,具体是基于BERT的语义交互模块,采用下文预测任务以学习文本间的句子关系特征,包括交互模块的编码层、交互模块的池化层和交互模块的归一化层;

[0058] 交互模块的编码层,将句子对送入BERT前,需要在句子的头部加入[CLS]标识符,并在两句之间插入[SEP]标识符进行切分。

将拼接好后的句子$T = \{[CLS], S_{A1}, S_{A2}, \ldots, S_{An}, [SEP], S_{B1}, S_{B2}, \ldots, S_{Bm}\}$送入BERT模型进行微调,输出$E_T = \{T_{[CLS]}, T_{A1}, T_{A2}, \ldots, T_{An}, T_{[SEP]}, T_{B1}, T_{B2}, \ldots, T_{Bm}\}$,即句子对的向量化表达;句子$T = \{[CLS], S_{A1}, S_{A2}, \ldots, S_{An}, [SEP], S_{B1}, S_{B2}, \ldots, S_{Bm}\}$具体是T = {[CLS], 好,好,学,习,[SEP], 天,天,向,上}。

[0059] 交互模块的池化层,通过BERT得到的句向量E_T通过池化层来提取重要特征缩小维度;

[0060] 交互模块的归一化层,句向量E_T经过层归一化后的输出结果为交互模块获取到的句子对的第二相似度表征向量。

[0061] 作为具体示例,采用社区问答数据集对模型进行训练,该数据集是一个大规模高质量的问答型数据集,该数据集针对某些社会问题进行提问,每个问题都有多个反馈回答,同一个问题的反馈可以作为相似舆情。

[0062] 对于交互模块的编码层,将句子对送入BERT前,需要在句子的头部加入[CLS]标识符,并在两句之间插入[SEP]标识符进行切分。将拼接好后的句子$T = \{[CLS], S_{A1}, S_{A2}, \ldots, S_{An}, [SEP], S_{B1}, S_{B2}, \ldots, S_{Bm}\}$送入BERT模型进行微调,输出$E_T = \{T_{[CLS]}, T_{A1}, T_{A2}, \ldots, T_{An}, T_{[SEP]}, T_{B1}, T_{B2}, \ldots, T_{Bm}\}$,即句子对的向量化表达。

[0063] 对于交互模块的池化层,通过BERT得到的句向量E_T通过池化层来提取重要特征缩小为度。平均池化主要用于当所有信息都应该有所贡献的时候,比如要获取全局上下文关系或者要获取网络深层的语义信息等。最大池化主要是为了减少无用信息造成的影响,同时它能降低特征维度并提取出更好、更强烈的语义信息特征。为了使模型的鲁棒性更强,这里用平均池化和最大池化共同处理特征向量即表征向量。句向量E_T平均池化后的结果为$E^T_{avg} = \sum_{i=0}^{n+m+2} \frac{E_{Ti}}{n}$,最大池化为$E^T_{max} = \max_{i=0}^{n+m+2} E_{Ti}$,其中$E^T_{avg}$是全局平均池化后获得的句子T的向量,$E^T_{max}$全局最大池化之后获得的句子T的向量。将平均池化的计算结果与最大池化的计算结果进行拼接,即$E^T_{pool} = Concat(E^T_{avg}, E^T_{max})$。

[0064] 对于交互模块的归一化层,E^T_{pool}经过层归一化后的输出结果为$E_T^{'} = LayerNorm(E^T_{pool})$。$E_T^{'}$即可作为交互模块的第二相似度表征向量。

[0065] 本申请的系统还包括匹配模块,利用此模块将孪生神经网络模块得到的第一相似度表征向量与基于BERT的交互模块得到的第二相似度表征向量拼接,得到句子A和句子B的最终相似度表征向量$E_{All} = Concat(E_{SNN}, E'_T)$。$E_{All}$既能表达句子对中实体词的之间的差异,又能通过结合BERT模型获取句子对的深层语义交互特征从而获得更加准确的文本相似度信息。最后通过SoftMax分类函数得到最终结果。

[0066] 在本发明的一种示例性实施例中,所述匹配模块中,具体SoftMax分类函数如下,$p(y = j)$代表的含义为样本向量 x 属于第 j 个分类的概率,其中W为权重系数,k表示有k个类别:

[0067] $$p(y = j) = \frac{e^{x^T W_j}}{\sum_{k=1}^{K} e^{x^T W_k}}$$

[0068] 将最终相似度表征向量E_{All}输入至softmax函数中,$E_{All} = Concat(E_{SNN}, E_T^{'})$,其中$E_{SNN}$为所述孪生神经网络模块的输出,$E'_T$为所述交互模块的输出,$E_{All}$为上述 softmax 函数中的 x;得到的最终结果p在[0,1]区间中,假设设置句子对A句子和B句子的文本相似的阈值为0.5,则当$p \geq 0.5$时,则认为A句子和B句子两文本匹配,否则两文本不匹配。

[0069] 实施例二

[0070] 使用实施例一中的基于孪生结构的舆情文本匹配的系统,本实施例提供一种基于孪生结构的舆情文本匹配的方法,包括如下步骤:

[0071] 构造孪生神经网络的编码层,利用孪生神经网络模块,提取出句子对中的命名实体信息,并对提取出的命名实体进行相似度计算,获取命名实体间的第一相似度表征向量;

[0072] 构造孪生神经网络的编码层,具体利用BERT+CRF模型(方法)构造孪生神经网络的编码层,包括两个相同或相似的神经网络建立的耦合三层架构,分别是输入层、特征提取层和相似度度量层,其中输入层输入需进行匹配的句子对,特征提取层将输入的句子对样本嵌入至高纬度空间得到句子对两个样本的表征向量,相似度度量层通过数学公式对提取出的两个样本的表征向量进行相似度计算,得到句子对的第一相似度表征向量。

[0073] 具体利用BERT+CRF模型(方法)构造孪生神经网络的编码层,从而提取出句子对中的命名实体信息,并对提取出的命名实体进行相似度计算,获取命名实体间的第一相似度表征向量,具体还包括:

[0074] 采用BERT层的掩码语言模型任务获取输入句子对语句中词级别的文本特征,再将BERT层输出的特征向量输入至CRF层;

[0075] 采用BERT层的下文预测任务,判断输入的句子对的A句子和B句子是否上下问相关,从而使模型学习到两个文本之间的关系,解决句子层面的问题;再将BERT层输出的特征向量输入至CRF层;

[0076] CRF层通过学习数据集中标签之间的转移概率,从而修正BERT层的输出;

[0077] 训练集即句子对采用BIO方法对实体进行标注,B(begin)表示该字符处于一个实体的开始,I(inside)表示该字符处于该实体的内部位置,O(outside)表示实体外部的不被关注的非实体字符;对于舆情文本需重点关注文本中的人名(PER)、地名(GEO)、以及组织(ORG),故训练集的实体标签有B-PER,I-PER,B-GEO,I-GEO,B-ORG,I-ORG,O这7种类型的标签;

[0078] 将句子对送入孪生神经网络前,需要在句子的头部加入[CLS]标识符,得到A、B句子对的A句子向量$T_A = \{[CLS], S_{A1}, S_{A2}, \ldots, S_{An}\}$和$T_B = \{[CLS], S_{B1}, S_{B2}, \ldots, S_{Bm}\}$;将$T_A$和$T_B$送入BERT进行微调,通过BERT层的编码为句子中每个位置上的字符引入上下文信息从而获取词性状态以进行表征向量,所有BERT的输出将作为CRF层的输入。

[0079] 获取句子对在语义方面的第二相似度表征向量,利用语义交互模块获取,具体包括:

[0080] 具体基于BERT采用下文预测任务以学习文本间的句子关系特征;

[0081] 将句子对送入BERT前,需要在句子的头部加入[CLS]标识符,并在两句之间插入[SEP]标识符进行切分。将拼接好后的句子$T = \{[CLS], S_{A1}, S_{A2}, \ldots, S_{An}, [SEP], S_{B1}, S_{B2}, \ldots, S_{Bm}\}$送入BERT模型进行微调,输出$E_T = \{T_{[CLS]}, T_{A1}, T_{A2}, \ldots, T_{An}, T_{[SEP]}, T_{B1}, T_{B2}, \ldots, T_{Bm}\}$,即句子对的向量化表达;

[0082] 通过BERT得到的句向量E_T通过池化层来提取重要特征缩小维度;

[0083] 句向量E_T经过层归一化后的输出结果为交互模块获取到的句子对的第二相似度表征向量。

[0084] 将第一相似度表征向量和第二相似度表征向量拼接,利用融合模块拼接,得到句子对的最终相似度表征向量;

[0085] 将最终相似度表征向量通过SoftMax分类函数得到文本匹配结果,利用匹配模块,将孪生神经网络模块得到的第一相似度表征向量与基于BERT的交互模块得到的第二相似度表征向量拼接,得到句子A和句子B的最终相似度表征向量$E_{All} = Concat(E_{SNN}, E_T^{'})$。$E_{All}$既能表达句子对中实体词的之间的差异,又能通过结合BERT模型获取句子对的深层语义交互特

征从而获得更加准确的文本相似度信息。最后通过SoftMax分类函数得到最终结果。总之，E_{SNN}可表达句子对中实体词的之间的差异，E'_T能获取句子对的深层语义交互特征，故可以获得更加准确的文本相似度信息。

[0086] 在本发明的一种示例性实施例中，所述匹配模块中，具体SoftMax分类函数如下，$p(y=j)$代表的含义为样本向量 x 属于第 j 个分类的概率，其中W为权重系数，k表示有k个类别：

[0087] $$p(y=j)=\frac{e^{x^T W_j}}{\sum_{k=1}^{K} e^{x^T W_k}}$$

[0088] 将最终相似度表征向量E_{All}输入至softmax函数中，$E_{All}=Concat(E_{SNN},E'_T)$，其中$E_{SNN}$为所述孪生神经网络模块的输出，$E'_T$为所述交互模块的输出，$E_{All}$为上述 softmax 函数中的 x；得到的最终结果p在[0,1]区间中，假设设置句子对A句子和B句子的文本相似的阈值为0.5，则当$p \geq 0.5$时，则认为A句子和B句子两文本匹配，否则两文本不匹配。

[0089] 为了进一步展示本发明的技术效果，将本发明提出的一种基于孪生结构的舆情文本匹配方法应用于STS-B语义相似度数据集。该数据集中每条数据包含句子对和相似度分数，相似度分数从0至5，分数越高则代表句子对的相似度越高，分数为0时则代表两个句子的语义不相似。且数据集被划分为训练集、验证集和测试集，其中，训练集中共有5231条数据，验证集包含1458条数据，测试集包含1361条数据。

[0090] 另外，为了更直观的进行对比，本发明同时利用文本匹配任务中的Siamese-CNN，Siamese-LSTM，ABCNN，BERT几个主流模型进行比较实验。最终不同模型在STS-B数据集上的实验结果如下所示：

[0091]

模型名称	模型准确率
Siamese-CNN	60.21
Siamese-LSTM	64.52
ABCNN	66.80
BERT	75.52
本发明提出的方法	83.96

[0092] 从以上实验结果可以发现孪生神经网络结构在语义相似度领域中应用时可以有效地提高模型的表现。该方法不仅通过句子对A句子和B句子两文本的语义特征进行相似性判断，还通过句子对A句子和B句子两文本的实体特征来判断文本是否是针对同一人物、事物或现象等等进行的描述，使文本数据相似度判断更加准确，从而使舆情文本系统的匹配准确率提升，减少了大量人力和时间进行人为地判断，提高了舆情文本分析的效率。

[0093] 实施例三

[0094] 另一方面，本发明还提供一种计算机可读的存储介质，所述计算机可读的存储介质包括存储的程序，其中，所述程序运行时执行上述的基于孪生结构的舆情文本匹配的方法。

[0095] 实施例四

[0096] 本发明还提供一种电子装置，包括存储器和处理器，所述存储器中存储有计算机程序，所述处理器被设置为通过所述计算机程序执行所述的基于孪生结构的舆情文本匹配的方法。

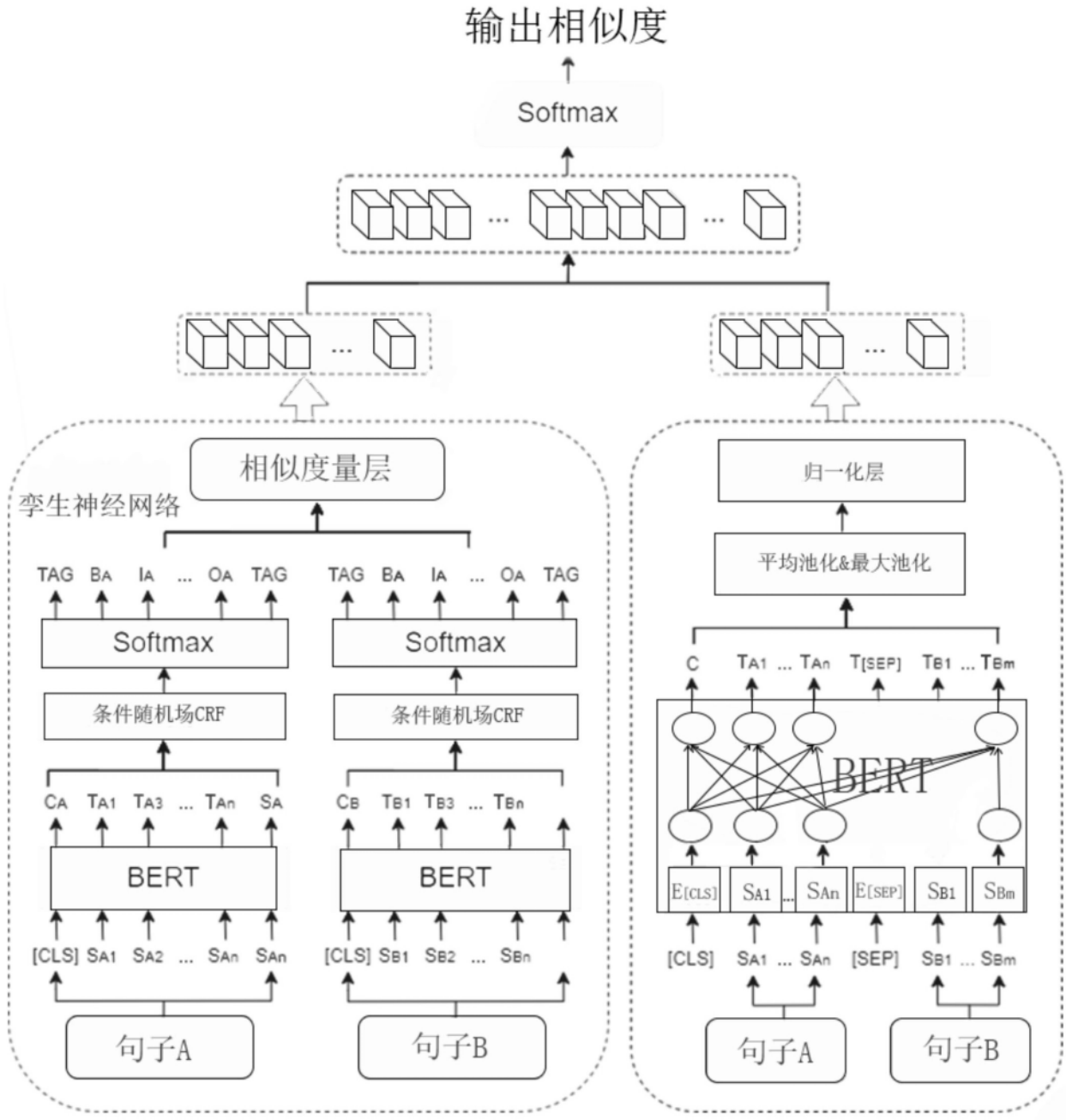

输出相似度
Softmax
相似度量层
孪生神经网络
TAG BA IA ... OA TAG
TAG BA IA ... OA TAG
Softmax
Softmax
条件随机场CRF
条件随机场CRF
CA TA1 TA3 ... TAn SA
CB TB1 TB3 ... TBn
BERT
BERT
[CLS] SA1 SA2 ... SAn SAn
[CLS] SB1 SB2 ... SBn
句子A
句子B
归一化层
平均池化&最大池化
C TA1 ... TAn T[SEP] TB1 ... TBm
BERT
E[CLS] SA1 ... SAn E[SEP] SB1 SBm
[CLS] SA1 ... SAn [SEP] SB1 ... SBm
句子A
句子B

图1

图2

字符	标签
张	B-PER
小	I-PER
红	I-PER
明	O
天	O
飞	O
北	B-GEO
京	I-GEO

图3