基于自注意力机制的事件时序关系分类方法

张义杰 李培峰 朱巧明
(苏州大学计算机科学与技术学院 江苏 苏州 215006)
(江苏省计算机信息处理技术重点实验室 江苏 苏州 215006)

摘 要 事件时序关系分类是事件抽取的重要后续任务。随着深度学习技术的发展,神经网络在事件时序关系分类任务中发挥着重要作用。但是,对于传统的循环神经网络或卷积神经网络而言,处理结构信息和捕获长距离依赖关系仍然是一个重大挑战。针对这个问题,文中提出了一种基于自注意力机制的事件时序关系分类模型架构,它可以直接捕获句子中任意两个词例之间的关系。将该机制与非线性网络层结合,可以使事件时序关系分类的性能得到显著提高。在TimeBank-Dense和Richer Event Description数据集上的对比实验证明:所提方法优于现有的大多数神经网络方法。

关键词 时序关系,深度学习,自注意力机制
中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.08.040

Event Temporal Relation Classification Method Based on Self-attention Mechanism

ZHANG Yi-jie LI Pei-feng ZHU Qiao-ming
(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China)
(Province Key Lab of Computer Information Processing Technology of Jiangsu, Suzhou, Jiangsu 215006, China)

Abstract Classifying temporal relation between events is a significant subsequent study of event extraction. With the development of deep learning, neural network plays a vital role in the task of event temporal relation classification. However, it remains a major challenge for conventional RNNs or CNNs to handle structural information and capture long distance dependence relations. To address this issue, this paper proposed a neural architecture for event temporal relation classification based on self-attention mechanism, which can directly capture relationships between two arbitrary tokens. The classification performance is improved significantly through combing this mechanism with nonlinear layers. The contrast experiments on TimeBank-Dense and Richer Event Description datasets prove that the proposed method outperforms most of the existing neural methods.

Keywords Temporal relation, Deep learning, Self-attention mechanism

1 引言

作为事件抽取(Event Extraction)的后续任务,事件时序关系分类被广泛应用于信息抽取、文档摘要、自动问答等自然语言处理任务中,且受到越来越多的关注,特别是在一系列TempEval[1)]评测任务之后。事件时序关系分类旨在对事件之间的时序关系(如"BEFORE""AFTER"等)进行正确分类。例如,在下面的句子中,事件"reinstated"和"convictions"的时序关系是"AFTER":

S1:A federal appeals court has **reinstated** his state **convictions** for securities fraud.

早期的研究工作[1-7]通常专注于提取文本中的词法、句法信息,或者借助各种外部知识库来获取语义信息,如WordNet和VerbOcean。但是,这些基于特征的方法严重依赖手工去构造特征和外部资源。此外,这些工作需要大量人工标注的实体属性特征,如事件类别(Class)、时态(Tense)、体态(Aspect)、极性(Polarity)等。因此,在实际应用场景中这些方法往往很难实现。

最近,在不依赖手工特征的情况下,深度学习框架的神经网络模型在事件时序关系分类任务中显示出优越性。在这些研究的基础上,本文提出了一种自注意力机制和非线性网络层相结合的模型架构。与RNNs和CNNs相比,自注意力机制有着自身的优势。首先,其输入位置和输出位置之间的依赖路径长度为1,而在RNNs中序列长度是n。由于RNNs将整个句子记忆编码至一个固定大小的向量,这就使得其对整个句子建模有时会成为一种负担,容易导致模型忽略关键的短语信息。而在自注意力机制中远距离的元素之间可以通过更短的路径进行交互,使得信息可以畅通无阻地通过网络;其

1) http://www.timeml.org/tempeval/

到稿日期:2018-07-09 返修日期:2018-10-28 本文受国家自然科学基金(61472265,61772354,61773276)资助。
张义杰(1994—),男,硕士生,CCF学生会员,主要研究方向为自然语言处理;**李培峰**(1971—),男,教授,博士生导师,CCF会员,主要研究方向为自然语言处理、机器学习,E-mail:pfli@suda.edu.cn(通信作者);**朱巧明**(1963—),男,教授,博士生导师,CCF会员,主要研究方向为中文信息处理。

次，与 CNNs 擅于抽取局部的位置不变特征相比，自注意力机制并不局限于固定的窗口尺寸大小。除此之外，自注意力机制使用加权求和的方法计算输出向量，在梯度传播方面要比 RNNs 或 CNNs 简单得多。因此，自注意力机制提供了一种更加灵活的方式来选择、表示和合成输入的信息，使其可以作为基于 RNNs 和 CNNs 的模型的良好补充。实验结果表明，将自注意力机制与非线性网络相结合的模型具有更好的分类能力与泛化性能。

2　相关工作

2.1　事件时序关系分类

传统的事件时序关系分类研究大多是基于模式匹配和统计机器学习方法。Marcu 等[4]提出了一种将单词配对作为有用特征的方法。Mani 等[5]在 TimeBank[1)] 标注语料的基础上利用事件属性构造特征向量，并使用最大熵分类器进行时序关系分类。Chambers 等[6]进一步结合了词性、句法树结构等语义特征，并从 WordNet 中提取词汇和形态学特征，使分类器性能得到提升。Li 等[7]挖掘了由事件语义衍生出的多种文档级约束条件，使用整数线性规划方法对分类器结果进行全局优化，进一步提升了系统性能。

随着深度学习的发展，各种神经网络模型被引入事件时序关系分类任务中。Cheng 等[8]和 Meng 等[9]分别将事件词之间的最短依存路径作为输入，构造基于 RNNs 的神经网络模型，在没有使用任何显性特征和外部资源的情况下，取得了十分可观的性能。Choubey 等[10]通过采用顺序学习能力良好的双通道 LSTM 来学习两个事件词上下文的句法和语义表示，提出了一种序列模型，用于相同句子间的事件时序关系分类。同样地，Tourille 等[11]提出了一种基于 LSTM 的神经网络架构，用于识别医学事件或时间表达式之间的包容关系。值得注意的是，由于依存分析对语句具有良好的表示性，因此以最短依存路径为输入的思想也被引入本文模型中。

2.2　自注意力机制

自注意力机制（Self-Attention）[12]已被成功地应用于多个 NLP 任务中。Cheng 等[13]使用 LSTMs 和自注意力机制来促进机器阅读任务。Lin 等[14]提出了基于自注意力的句子级嵌入技术，并将它们应用到识别特征提取、情感分析和文本蕴含关系等任务中。Paulus 等[15]结合强化学习与自注意力机制来捕捉抽象式摘要任务中的长距离依赖关系。Vaswani 等[12]将自注意力机制应用于神经机器翻译，取得了目前最优的结果。最近，Shen 等[16]将自注意力机制引入语言理解任务中，并在多个数据集上都取得了最优性能。本文提出的方法遵循这条研究线路，首次把自注意力机制引入时序分类任务，通过自注意力机制学习长距离依赖关系。实验结果证明了自注意力机制在事件时序关系分类任务中的有效性。

3　基于自注意力机制的事件时序分类模型

图 1 给出了基于自注意力机制的事件时序分类模型的网络架构图。该架构主要由左、右两个相同的网络模块组成，这两个网络模块分别对应输入层中左、右两个最短依存路径序列。其中每一个模块都包含一个非线性子层，其后跟随一个自注意力子层。模型的顶端是用于关系分类的 Softmax 层，其输入为左、右两个分支的输出拼接而成的向量。

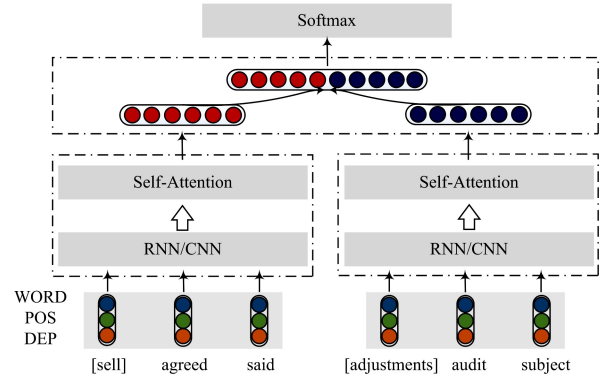

图 1　基于 Self-attention 的事件时序关系分类模型

Fig. 1　Event temporal relation classification model based on Self-attention mechanism

3.1　最短依存路径

最短依存路径（SDP）已多次在关系抽取的任务中体现出对于捕获两个实体之间关系结构的有效性，并且已被证明在此任务中对于识别事件时序关系具有巨大潜力。给定一个事件对，根据其所在事件句的依存解析树提取最短依存路径。对于处于同一句子中的事件对，它们之间的依存路径被对应结点的公共祖先结点分割成两个最短依存路径分支；对于处于相邻句子中的事件对，本文沿用 Cheng 等[8]的假设，认为两个相邻句子的依存解析树共享同一个“根结点”，这样相邻句的依存路径也可以用同样的方法表示。例如，图 2 给出的是两个相邻句子“The company said it has agreed to sell the extrusion division.”和“The sale of the extrusion division is subject to audit adjustments for working capital changes through the closing.”对应的依存解析树及最短依存路径表示，其中红色箭头即是事件对（sell，adjustments）之间的最短依存路径。

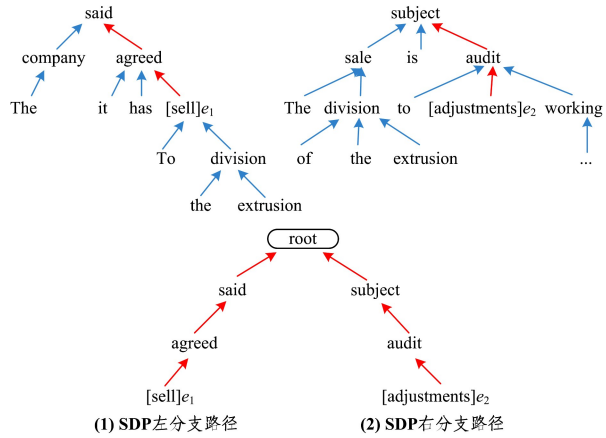

图 2　依存解析树和最短依存路径（电子版为彩色）

Fig. 2　Dependency parse tree and SDP

对于一条 SDP 分支中的每一个单词，本文首先将词本身

1) https://catalog.ldc.upenn.edu/LDC2006T08

(W)、词性(P)和该词结点在依存解析树中与其父结点的依存关系(D)映射到低维实值向量空间中。其中词本身通过使用 Word2Vec 预训练词向量的方式进行查表映射,词性与依存关系通过随机初始化的方法进行映射。之后分别获得词向量 $\boldsymbol{x}_w$、词性向量 $\boldsymbol{x}_p$ 和依存向量 $\boldsymbol{x}_d$,将这 3 个向量进行拼接形成每个单词的向量表示 $\boldsymbol{X}_i$:

$$\boldsymbol{x}_i=\boldsymbol{x}_w \oplus \boldsymbol{x}_p \oplus \boldsymbol{x}_d \tag{1}$$

最终得到两个分支序列的向量化表示 $\boldsymbol{S}_1=\{\boldsymbol{X}_{11},\boldsymbol{X}_{12},\boldsymbol{X}_{13},\cdots\}$ 和 $\boldsymbol{S}_2=\{\boldsymbol{X}_{21},\boldsymbol{X}_{22},\boldsymbol{X}_{23},\cdots\}$。

3.2 自注意力机制

自注意力,也称作内部注意力,是一种特殊的注意力机制,它只需要序列本身就可以计算其表示。该机制先通过 3 个不同的线性变换将输入向量矩阵 $\boldsymbol{X}\in\mathbb{R}^{n\times d}$ 映射成 queries 矩阵 $\boldsymbol{Q}\in\mathbb{R}^{n\times d}$、keys 矩阵 $\boldsymbol{K}\in\mathbb{R}^{n\times d}$ 和 values 矩阵 $\boldsymbol{V}\in\mathbb{R}^{n\times d}$,再利用下式计算注意力得分:

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{2}$$

其中,d 是网络中隐藏单位的数量。

本文引入了 Vaswani 等[12]提出的多头注意力(Multi-Head Attention)结构。首先通过不同的平行头计算 queries 与 keys 的相关性,然后将所有平行头的输出结果拼接成一个单独的向量。计算公式如下:

$$\boldsymbol{M}_i=\text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^{Q},\boldsymbol{K}\boldsymbol{W}_i^{K},\boldsymbol{V}\boldsymbol{W}_i^{V}) \tag{3}$$

$$\boldsymbol{M}=\text{Concat}(\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_h) \tag{4}$$

$$\boldsymbol{Y}=\boldsymbol{M}\boldsymbol{W} \tag{5}$$

其中,h 为平行头的数量。$\boldsymbol{W}_i^{Q}\in\mathbb{R}^{n\times d/h}$,$\boldsymbol{W}_i^{K}\in\mathbb{R}^{n\times d/h}$,$\boldsymbol{W}_i^{V}\in\mathbb{R}^{n\times d/h}$ 和 $\boldsymbol{W}\in\mathbb{R}^{d\times d}$ 是对应线性变换的权重矩阵。

3.3 非线性子层

神经网络的成功源于其高度灵活的非线性转换能力。由于自注意力机制使用加权求和的方法来计算输出向量,因此它对语句特征的表示能力相对有限。为了进一步提高注意力网络的表示能力,本文使用一个非线性子层来对底层输入进行转换。本文主要使用了目前主流的两种非线性神经网络,即循环神经网络(RNN)和卷积神经网络(CNN)。

3.3.1 循环神经网络

使用双通道 GRU[17]来构建模型的循环子层。GRU 是 RNN 的改进版本,与另一种 RNN 变体 LSTM 相比,它保持了 LSTM 的效果的同时,使网络结构变得更为简单,从而提高了模型的训练速度。给定一个输入向量序列 $\{\boldsymbol{x}_t\}$,两个 GRU 分别在相反的方向对输入进行编码。为了在输入和输出之间保持相同的向量维度,本文使用按位求和操作来组合两个向量表示:

$$\overrightarrow{\boldsymbol{h}_t}=\text{GRU}(\boldsymbol{x}_t,\overrightarrow{\boldsymbol{h}_{t-1}}) \tag{6}$$

$$\overleftarrow{\boldsymbol{h}_t}=\text{GRU}(\boldsymbol{x}_t,\overleftarrow{\boldsymbol{h}_{t-1}}) \tag{7}$$

$$\boldsymbol{y}_t=\overrightarrow{\boldsymbol{h}_t}+\overleftarrow{\boldsymbol{h}_t} \tag{8}$$

3.3.2 卷积神经网络

对于卷积子层,本文采用 Dauphin 等[18]提出的 GLU (Gated Linear Unit)模型。与标准的 CNN 相比,该模型在多个自然语言处理任务中被证明更容易进行学习且效果更优。给定两个滤波器 $\boldsymbol{W}\in\mathbb{R}^{k\times d\times d}$ 和 $\boldsymbol{V}\in\mathbb{R}^{k\times d\times d}$,GLU 的输出计算如下:

$$GLU(\boldsymbol{X})=(\boldsymbol{X}*\boldsymbol{W})\odot\sigma(\boldsymbol{X}*\boldsymbol{V}) \tag{9}$$

其中,σ 为激活函数。

3.4 关系预测和模型训练

采用 Softmax 层作为输出层来获得关系标签的预测结果:

$$\boldsymbol{Y}=\text{Concat}(\boldsymbol{Y}_1+\boldsymbol{Y}_2) \tag{10}$$

$$\boldsymbol{o}=\text{Softmax}(\boldsymbol{Y}\boldsymbol{W}_o+\boldsymbol{b}_o) \tag{11}$$

其中,$\boldsymbol{Y}_1$ 和 $\boldsymbol{Y}_2$ 分别对应左、右两个分支序列经过非线性层和自注意力机制后的输出。$\boldsymbol{W}_o$ 和 $\boldsymbol{b}_o$ 是分别表示 Softmax 函数的权重矩阵和偏置量。

为了获取最优模型,本文采用随机梯度下降算法来最小化负对数似然函数。目标函数定义为:

$$J(\theta)=-\sum_{i=1}^{T}\log p(y_i\mid\boldsymbol{x}_i,\theta) \tag{12}$$

其中,θ 为模型可训练参数集合;T 表示训练样本个数;$(\boldsymbol{x}_i,y_i)$表示训练样本中第 i 个样本 $\boldsymbol{x}_i$ 对应的标签关系是 y_i。

4 实验与分析

4.1 数据集及评价指标

本文在 TimeBank-Dense(TB-D)与 Richer Event Description(RED)[1)]两个数据集上验证所提方法。TB-D 的标注机制旨在缓解原始 TimeBank 语料的样例稀疏问题。与 183 个原文档只包含 6418 个事件对相比,TB-D 可以在 36 个文档中标注 12715 个事件对。RED 语料库取材于 95 篇英文新闻专线、研讨论坛和叙述性文本文档,共包含 4209 个事件对。它定义了 11 种时序关系类型,其中 4 种还包含了额外的因果关系信息。本文通过移除关系类型中的因果信息,并且把“CONTAINS”类型和“CONTAINS-SUBEVENT”类型融合成一种,将 11 种时序关系类型减少到 6 种。两种数据集的具体关系类型如表 1 所列。其中,TB-D 语料中的“VAGUE”类型旨在处理模糊的时序关系,或表示无明确时序关系存在的事件对。

表 1 TB-D 和 RED 关系类型

Table 1 Relation type of TB-D and RED

TB-D	RED
AFTER	BEFORE
BEFORE	CONTAINS
SIMULTANEOUS	OVERLAP
INCLUDES	BEGINS-ON
IS_INCLUDED	ENDS-ON
VAGUE	SIMULTANEOUS

为了评价本文的方法的有效性,采用准确率(Precision)、召回率(Recall)以及 $F1$ 值作为实验性能的评价指标。

4.2 参数设置

为了与已有工作进行公平的比较,本实验采用句子级的 5 倍交叉验证方式进行模型调优,取微平均值作为最终结果,其中验证集从训练集中以 15% 的比例随机抽样获取。为了防止模型过拟合,采用 L2 正则化方法对网络参数进行约束,

1) https://catalog.ldc.upenn.edu/LDC2016T23

训练过程中在非线性子层(RNN 或 CNN)引入 dropout 策略，并使用早停法(early stopping)来保存验证集上的最好模型。另外，本文使用 Adam 优化方法进行参数更新，其他的参数设置如表 2 所列。

表 2　参数设置

Table 2　Settings of hyperparameters

词向量维度	200
词性向量维度	50
依存关系向量维度	50
GRU 单元大小	230
CNN 滤波器宽度	3
Dropout 比率	0.5
Adam 学习率	10^{-3}
批大小	64
Epoch 数	20
L2 惩罚权重系数	10^{-5}

4.3　实验结果与分析

4.3.1　与传统基于特征的方法比较

将以下两种性能优异的传统方法作为基准系统。1)CAEVO：该方法 Chambers 等[6]于 2014 年提出的一种基于滤网架构的混合模型，结合了若干个基于规则和统计机器学习的分类器；它在每个分类器预测标签之后还添加了用于过渡的推理机制。2)MIRZA：该方法于 2016 年由 Mirza 等[19]提出，使用 L2 正则化的逻辑斯特回归构建分类器，并将低维度的词向量与稀疏的传统特征向量拼接作为分类器的输入，其中，分类器使用的特征包括实体属性、时序信号词和 WordNet 语义信息等。

表 3 列出了在 TB-D 数据集上具体关系类型和总体的实验结果对比，其中 CAEVO 只给出了总体的 $F1$ 值。从表中可以看到，基于自注意力机制的神经网络模型在所有关系类型和总体表现上都胜过传统的基于特征的方法。其中，采用 RNN(GRU)作为非线性子层的网络模型在总体的 $F1$ 值上相比 MIRZA 提升了 2.6%；并且，在两个样例数最多的 AFTER 和 BEFORE 关系类型上，分别提高了 6.3%和 4.7%。实验结果证明了神经网络方法与基于人工设计的特征与规则的方法相比，在挖掘事件句隐含的深层语义信息上的优越性。

表 3　在 TB-D 数据集上与传统基于特征的方法的 F1 值比较

Table 3　F1 comparison with traditional feature-based methods on TB-D dataset

Relation	CAEVO	MIRZA	Self-At (RNN)	Self-At (CNN)
AFTER	—	43.0	**49.3**	48.4
BEFORE	—	47.1	**51.8**	50.2
SIMULTANEOUS	—	—	—	—
INCLUDES	—	4.9	**9.0**	8.6
IS_INCL	—	25.0	**31.2**	28.3
VAGUEUDED	—	61.3	**63.0**	62.8
Overall	49.4	51.9	**54.5**	53.8

另外，在采用不同线性子层的对比上，从实验结果可以看出，RNN 相比 CNN 在性能表现上更具优势。原因可能在于 CNN 在捕获句子特征的过程中只能抽取位置不变的特征，缺乏对全局上下文信息的考虑；与 CNN 不同的是，RNN 中的记忆模块使其能够充分学习整个依存路径的序列信息，因此更加适合自然语言处理任务。值得注意的是，由于 SIMULTANEOUS 关系类型的样例数过于稀少，导致传统方法与神经网络方法都无法正确识别。

4.3.2　自注意力机制对实验性能的影响

为了验证模型中自注意力机制的引入对时序关系分类任务带来的影响，本文设置了 3 个基准模型，并分别在 TB-D 和 RED 两个数据集上进行实验。第一个模型是 Cheng 等[8]于 2017 年提出的基于双通道 LSTM 的分类模型；第二个模型与第三个模型分别只采用了 RNN 和 CNN 单层神经网络，并没有引入自注意力机制。3 个基准模型都采用最短依存路径作为输入，并同样将路径中每个词的词本身、词性以及依存关系作为输入特征。对比实验结果如表 4 和表 5 所列。

表 4　在 TB-D 数据集上的对比

Table 4　Comparison on TB-D dataset

Relation	LSTM	GRU	CNN	Self-At (GRU)	Self-At (CNN)
AFTER	44.0	43.7	43.0	**49.3**	48.4
BEFORE	46.0	46.1	44.5	**51.8**	50.2
SIMULTANEOUS	—	—	—	—	—
INCLUDES	2.5	4.5	3.0	**9.0**	8.6
IS_INCL. UDED	17.0	16.2	16.0	**31.2**	28.3
VAGUE	62.4	62.3	62.0	**63.0**	62.8
Overall	52.9	52.8	52.3	**54.5**	53.8

表 5　在 RED 数据集上的对比

Table 5　Comparison on RED dataset

Relation	LSTM	GRU	CNN	Self-At (GRU)	Self-At (CNN)
BEFORE	61.5	61.6	61.2	**63.9**	63.5
CONTAINS	55.5	55.2	55.0	**60.1**	59.3
OVERLAP	32.6	32.5	32.6	**34.4**	34.0
BEGINS-ON	22.1	21.9	21.8	**22.3**	22.2
ENDS-ON	19.3	19.5	19.0	20.1	**20.2**
SIMULTANEOUS	—	—	—	—	—
Overall	51.5	51.4	51.0	**55.6**	54.9

首先，从 TB-D 和 RED 两个数据集的实验结果中可以看出，在不引入自注意力机制的情况下，基于 GRU 的 RNN 模型可以取得与 LSTM 相近的性能表现，但在模型训练速度上更具优势。当引入自注意力机制之后，Self-At(GRU)方法和 Self-At(CNN)方法比只采用单层非线性层的效果更优。原因在于自注意力机制能够有效学习非线性层的输出序列内部的长距离依赖关系，因此在学习过程中可以捕获更多的上下文语义信息。并且，多头结构的引入使得模型能够从不同的表示子空间里学习更多的相关信息。以下是 TB-D 数据集中的一个例子。

S2：Ms. Sanders was hit several times and was pronounced dead at the scene.

S3：The other customers fled，and the police said it did not appear that anyone else was injured.

其中，事件对(dead，said)的真实时序标签为 AFTER，基于单层 LSTM/GRU/CNN 的模型都将其错误地预测为 INCLUDES 类型或 VAGUE 类型。而自注意力机制的引入使得模型在对“pronounced”“police”等单词通过非线性层编码之后学习到更多依赖信息，从而充分挖掘其与事件词之间的上下文联系，最终正确预测该事件对的时序关系。

结束语　本文提出了一种将自注意力机制和非线性网络

层相结合的模型,用于事件时序关系分类任务。该模型克服了传统的神经网络模型无法有效处理文本的结构信息以及捕获句子长距离依赖关系的问题,使得实验性能明显提升。但是自注意力机制对于序列顺序信息的捕捉依然存在缺陷。下一步工作将着重考虑如何在模型中加入词的位置信息来捕捉序列顺序。

参 考 文 献

[1] LIN J,YUAN C F. Extraction and Computation of Chinese Temporal Relation[J]. Journal of Chinese Information Processing,2009,23(5):62-67. (in Chinese)
林静,苑春法. 汉语时间关系抽取与计算[J]. 中文信息学报,2009,23(5):62-67.

[2] ZHONG Z M,LIU Z T,ZHOU W,et al. The Model of Event Relation Representation[J]. Journal of Chinese Information Processing,2009,23(6):56-60. (in Chinese)
仲兆满,刘宗田,周文,等. 事件关系表示模型[J]. 中文信息学报,2009,23(6):56-60.

[3] WANG F E,TAN H Y,QIAN Y L. Recognition of Temporal Relation in One Sentence Based on Maximum Entropy[J]. Computer Engineering,2012,38(4):37-39. (in Chinese)
王风娥,谭红叶,钱揖丽. 基于最大熵的句内时间关系识别[J]. 计算机工程,2012,38(4):37-39.

[4] MARCU D,ECHIHABI A. Anunsupervised approach to recognizing discourse relations[C]// Proceedings of the Association for Computational Linguistics. Association for Computational Linguistics,2002:368-375.

[5] MANI I,VERHAGEN M,WELLNER B,et al. Machine learning of temporal relations[C]// Proceedings of the Association for Computational Linguistics. Association for Computational Linguistics,2006:753-760.

[6] CHAMBERS N,WANG S,JURAFSKY D. Classifying temporal relations between events[C]// Proceeding of the ACL on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics,2007:173-176.

[7] LI P F,ZHU Q M,ZHOU G D,et al. Global Inference to Chinese Temporal Relation Extraction[C]// Proceedings of the International Conference on Computational Linguistics. 2016:1451-1460.

[8] CHENG F,MIYAO Y. Classifying Temporal Relations by Bidirectional LSTM over Dependency Paths[C]// Proceedings of the Association for Computational Linguistics(Short Papers). Association for Computational Linguistics. 2017:1-6.

[9] MENG Y,RUMSHISKY A,ROMANOV A. Temporal Information Extraction for Question Answering Using Syntactic Dependencies in an LSTM-based Architecture[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics,2017:887-896.

[10] CHOUBEY P K,HUANG R H. A Sequential Model for Classifying Temporal Relations between Intra-Sentence Events[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2017:1796-1802.

[11] TOURILLE J,FERRET O,TANNIER X,et al. Neural Architecture for Temporal Relation Extraction:A Bi-LSTM Approach for Detecting Narrative Containers[C]// Proceedings of the Association for Computational Linguistics. Association for Computational Linguistics,2017:224-230.

[12] VASWANI A,SHAZEER N,PARMAR N,et al. Attentionis all you need[J]. arXiv:1706.03762.

[13] CHENG J P,DONG L,LAPATA M. Long Short-Term Memory-Networks for Machine Reading[J]. arXiv:1601.06733.

[14] LIN Z H,FENG M W,SANTOS C N,et al. A Structured Self-attentive Sentence Embedding[J]. arXiv:1703.03130.

[15] PAULUS R,XIONG C M,SOCHER R. A Deep Reinforced Model for Abstractive Summarization[J]. arXiv:1705.04304.

[16] SHEN T,ZHOU T Y,LONG G D,et al. DiSAN:Directional Self-Attention Network for RNN/CNN-free Language Understanding[J]. arXiv :1709.04696.

[17] DEY R,SALEMT F M. Gate-variants of Gated Recurrent Unit (GRU)neural networks[J]. arXiv:1701.05923.

[18] DAUPHIN Y N,FAN A,AULI M,et al. Language Modeling with Gated Convolutional Networks[J]. arXiv:1612.08083.

[19] MIRZA P,TONELLI S. On the contribution of word embeddings to temporal relationclassification[C]// Proceedings of the International Conference on Computational Linguistics. 2016:2818-2828.