Computer Science and Application 计算机科学与应用, 2022, 12(2), 405-411
Published Online February 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.122041

Hans汉斯

中文突发事件论元识别方法研究

李明亮，王　勇，王　瑛

广东工业大学计算机学院，广东　广州

收稿日期：2022年1月18日；录用日期：2022年2月15日；发布日期：2022年2月22日

摘　要

事件抽取是从非结构化的文本中抽取用户感兴趣的事件信息，并以结构化的形式展现。当前社交媒体快速发展，互联网上突发事件信息数量也剧增。如何准确地从大量无结构事件信息中识别并抽取出突发事件信息，分析突发事件舆情趋势对社会安全极为重要。本文利用双向长短期记忆神经网络(BiLSTM)与图注意力机制(GAT)获取事件句子中的句法信息和获得事件论元之间的内在关联，进一步提升事件论元识别的准确率。通过在真实数据集中多次实验证明，本文中的方法在公开的数据集上进行验证，与以往的事件论元识别方法相比获得较大的性能提升。

关键词

事件论元识别，双向长短期记忆神经网络，图注意力机制

Research on the Method of Chinese Emergency Event Argument Recognition

Mingliang Li, Yong Wang, Ying Wang

School of Computer Science and Technology, Guangdong University of Technology, Guangzhou Guangdong

Received: Jan. 18th, 2022; accepted: Feb. 15th, 2022; published: Feb. 22nd, 2022

Abstract

Event extraction is to extract the event information that users are interested in from unstructured text and display it in a structured form. With the rapid development of social media, the amount of emergency event information on the Internet has also increased dramatically. How to accurately identify and extract emergent event information from a large amount of unstructured event information, and analyze the trend of public opinion on emergencies is extremely important to social security. In this paper, the Bi-directional Long Short-Term Memory (BiLSTM) and the graph attention network (GAT) are used to obtain the syntactic information in the event sentence and

文章引用: 李明亮, 王勇, 王瑛. 中文突发事件论元识别方法研究[J]. 计算机科学与应用, 2022, 12(2): 405-411.
DOI: 10.12677/csa.2022.122041

obtain the intrinsic correlation between the event arguments, so as to further improve the accuracy of the event argument recognition. Multiple experiments on real datasets prove that the method in this paper is validated on public datasets and achieves a large performance improvement compared with previous event argument extraction methods.

Keywords

Event Argument Recognition, BiSTM, Graph Attention Network

1. 引言

突发事件，是指突然发生，造成或者可能造成严重社会危害，需要采取应急处置措施予以应对的自然灾害、事故灾难、公共卫生事件和社会安全事件。突发事件的发生不仅会扰乱社会秩序，而且损坏人民群众生命财产，对社会造成严重危害的同时对公安维护社会稳定工作和保护人民财产安全提出了严峻考验。伴随社交媒体快速发展，互联网上突发事件信息数量也剧增。如何准确地从大量无结构事件信息中识别并抽取出突发事件信息，分析突发事件舆情趋势对社会安全极为重要。

ACE 认为事件是事物状态的改变或事情的发生，将事件抽取定义为从非结构化的文本中识别并抽取出事件信息，并以结构化的形式展现，包括事件的触发词、事件类型、事件论元、论元角色[1]。事件抽取对人们认知世界有着深远意义和重大应用价值，也是信息检索、智能问答、知识图谱构建等实际应用的基础，是目前自然语言处理领域中比较突出的研究热点之一[2]。事件抽取可以分为事件识别和事件论元识别两个部分，事件识别是指抽取对应的触发词并确定事件所属类型，事件论元识别是指抽取事件中包含的论元并确定论元所对应的角色。本文的研究内容是突发事件论元识别，研究目的是找出突发事件论元在事件句中的位置，并以结构化地展示。

2. 相关工作

传统的事件抽取任务，分别是基于结构分析的模式匹配方法和基于数理统计的机器学习方法。Ahn 等[3]首先对事件抽取进行了定义，将事件抽取任务当作多分类任务，使用词级特征、句级特征以及外部知识对触发词和论元进行分类。McClosky 等[4]将事件抽取任务看成依存分析问题，将事件触发词与论元之间的关系树看成依存关系树，以此提升模型性能。Li 等[5]提出了基于结构化感知器的联合模型，同时抽取触发词和论元。该工作中还设计了如触发词和论元的词性、语法、语义等局部特征和能够进行触发词和论元交互的全局特征帮助模型进行分类。

基于神经网络的事件抽取方法与传统的事件抽取方法相比，不仅避免了大量的人工标注成本以及复杂的特征工程，而且泛华能力强，目前已在事件抽取领域获得了广泛的应用[6]。Chen 等人[7]首次将神经网络模型应用到事件抽取领域，该文章提出一种基于动态多池的神经网络模型，在不借助复杂的自然语言处理工具情况下，能够捕获句子级别的信息，并且动态的保留多个句子信息。Sha 等人[8]提出一种 BiLSTM 的神经网络模型，该模型利用 BiLSTM 获取句子级别的特征，并利用 CNN 获取句子中的局部上下文信息，该模型无需任何人为地提供特征。Lin [9]等人在对每个单词建模时使用了依赖桥来增强单词

之间的相互关系，并利用张量层同时捕获事件论元之间的关系以及其在事件的角色。Yang [10]等人提出了一个基于预训练语言模型的框架，该框架包含一个作为基础的事件抽取模型以及一种生成被标记事件的方法。我们提出的事件抽取模型由触发词抽取器和论元抽取器组成，论元抽取器用前者的结果进行推理。此外，我们根据角色的重要性对损失函数重新进行加权，从而提高了论元抽取器的性能。Zheng 等人[11]提出了 Doc2EDGA 模型，定义了一个事件填充框架，先对事件进行检测，再按照一定的顺序进行论元解码，解决了部分论元重叠的问题。

在以往的事件论元识别方法中，往往只关注句子之间的线性信息，而忽略了事件论元之间存在的相互依存的关系，这样会导致事件论元识别的准确度降低。本文将针对上述问题，提出利用双向长短期记忆神经网络(BiLSTM)获取句子中的上下文语义信息，并通过图注意力机制(GAT)获取事件论元之间的依存关系，最后在公开的数据集上进行对比实验，以验证方法的可行性。

3. 本文方法

本文提出的的 BiLSTM-GAT 模型由词嵌入层，BiLSTM 层，GAT 层和输出层组成，总体架构如图 1 所示。

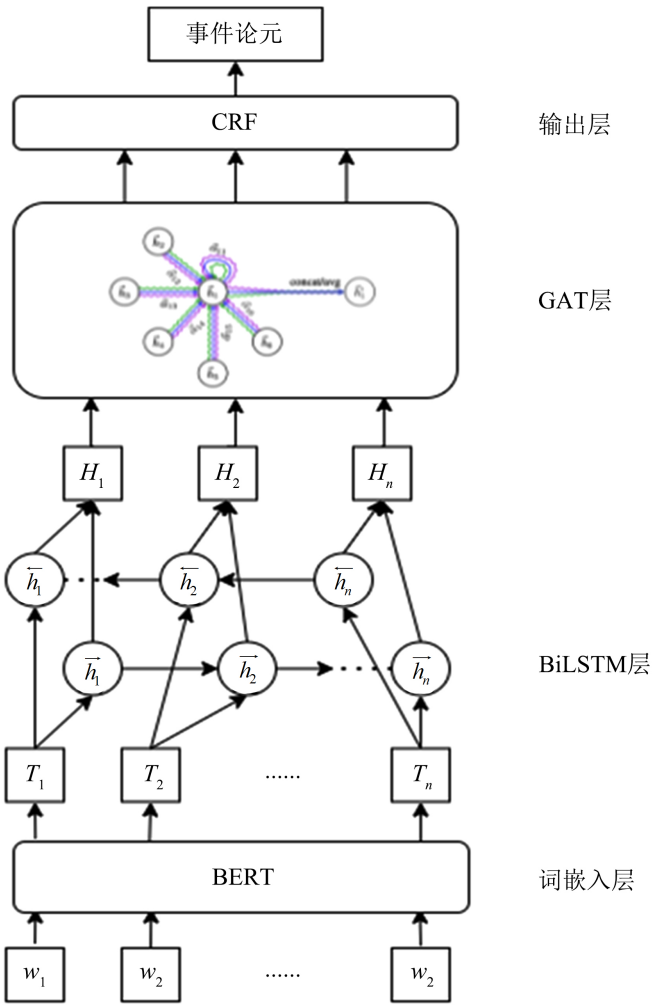

Figure 1. Overall architecture diagram of BiLSTM-GAT model
图 1. BiLSTM-GAT 总体架构图

3.1. 任务定义

将论元抽取定义为序列标注任务，采用 BIO 序列标注法进行标注，及将每个元素表示为“B-X”，“I-X”，“O”。其中“B-X”标签表示此元素所在的标注序列中属于 X 类型并且是这个标注序列的开头，“I-X”表示此元素所在的标注序列中属于 X 类型的并且是这个标注序列的中间字段，O 表示的是标注序列中的其它字段。将事件论元的标签类型定义为 P，T，L，其中 P (Participant)代表事件的参与者，T (Time)代表事件发生的时间，L (Location)代表事件发生的地点。

3.2. 词嵌入层

本次实验使用 BERT 中文预训练模型作为词嵌入层。BERT 中文预训练模型采用双向 Transformer 搭建，在维基百科中文语料集采用全词覆盖以及语法桥的方式进行训练，能够捕获词语间的表示以及句子级的表示，在许多自然语言处理任务上表现出色。

将输入句子经过 WordPiece 切分后，在句子开始位置加入特殊标记[CLS]，在句子末尾加入特殊标记[SEP]，由此，给定输入序列 = {1, 2, ···,}，n 代表序列的长度，而 BERT 的输入比较特别，包括三部分，Token Embedding：即传统意义的词向量，每一个 Token 对应一个向量；Segment Embeddings：标记输入的 Token 是属于句子 A 还是句子 B；Position Embeddings：具体标记每一个 Token 的位置。将三个向量对应相加得到 BERT 的输入向量 $E=\{e_1,e_2,\cdots,e_n\}$，经过 BERT 编码后得到 $T=\{h_1^{bert},h_2^{bert},\cdots,h_n^{bert}\}$。

3.3. BiLSTM 层

BiLSTM 在设计上可以很好的捕捉句子之间的长距离依赖关系，在对文本进行特征提取时，充分考虑到了文本前后文信息之间的相互影响。通过 BiLSTM 层进行编码可以对句子进行从前到后以及从后到前的完整的上下文信息保存。

BiLSTM 模型是由 t 时刻的输入词 X_t，细胞状态 C_t，临时细胞状态 $\tilde{C}_t$，隐层状态 h_t，遗忘门 f_t，记忆门 i_t，输出门 o_t 组成。通过对细胞状态中信息遗忘和记忆新的信息使得对后续时刻计算有用的信息得以传递，而无用的信息将被丢弃，并在每个时间步都会输出隐藏层状态 h_t。将 BERT 编码得到的 T 输入到 BiLSTM 层，如公式(1)所示

$$h_t=\left[\overrightarrow{LSTM}\left(h_t^{bert}\right);\overleftarrow{LSTM}\left(h_t^{bert}\right)\right] \tag{1}$$

3.4. GAT 层

GAT 在图卷积网络的基础上加入注意力机制去计算每个节点的邻居节点对它的权重，使得从局部信息可以获取到整个网络整体信息却无需提前知道整个网络的结构，从而学习到突发事件论元之间的相互依存关系。同时，GAT 通过堆叠这些隐藏自注意层能够获取临近点的特征，从而避免大量矩阵运算，使得计算更加高效。将 BiLSTM 隐层输入到 GAT 中，输入是一组顶点特征 $h=\{\vec{h_1},\vec{h_2},\cdots,\vec{h_n}\}$，$\vec{h_i}\in R^F$，其中 N 是顶点数，F 是每个顶点的特征数。这个层会生成一组新的顶点特征，$h'=\{\vec{h_1'},\vec{h_2'},\cdots,\vec{h_n'}\}$，$\vec{h_i'}\in R^{F'}$ 作为输出。

将每个顶点使用一个共享参数的线性变换，参数为 $W\in R^{F\times F'}$，然后在每个顶点做一个自注意力机制共享运算 $a=R^{F'}\times R^{F'}\to R$：

$$e_{ij}=a\left(W\vec{h_i},W\vec{h_i}\right) \tag{2}$$

为了使得注意力系数更容易计算和便于比较，引入 softmax 对所有的 i 的相邻节点 j 进行正则化：

$$\alpha_{ij}=\text{sofxmax}_j\left(e_{ij}\right)=\frac{\exp e_{ij}}{\sum_{k\in N_i}\exp e_{ik}} \tag{3}$$

综合上述(2)和(3)公式，整理得到完整的注意力机制如下：

$$\alpha_{ij}=\frac{\exp\left(\text{Leaky ReLu}\left(\vec{a}^T\left[W\vec{h}_i\middle\|W\vec{h}_j\right]\right)\right)}{\sum_{k\in N_i}\exp\left(\text{Leaky ReLu}\left(\vec{a}^T\left[W\vec{h}_i\middle\|W\vec{h}_j\right]\right)\right)} \tag{4}$$

通过上述运算得到了正则化后的不同节点之间的注意力系数，可以用来预测每个节点的输出特征：

$$\vec{h}_i'=\sigma\left(\sum_{j\in N_i}\alpha_{ij}W\vec{h}_j\right) \tag{5}$$

为了稳定自注意力机制学习过程，使用 K 个独立的注意力机制执行公式(4)，然后采用 K 平均连接，并延迟应用最终的线性函数，得到最终公式：

$$\vec{h}_i'=\sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N_i}\alpha_{ij}^kW^k\vec{h}_j\right) \tag{6}$$

3.5. 输出层

条件随机场是一种序列化标注算法，用于解决在给定一组输人随机序列的情况下，预测另一组输出随机序列的概率分布。本文使用 CRF 作为解码器，CRF 通过 GAT 的输入参数学习序列标签的路径分布，其概率计算公式如下：

$$L(W,b)=\sum_i\log p\left(y\middle|h;W,b\right) \tag{7}$$

最后获得标注序列时，概率最大的候选标注序列就是最终所需结果，公式为：

$$y^*=\arg\max_{y\in Y(h)}p\left(y\middle|h;W,b\right) \tag{8}$$

4. 实验

4.1. 实验环境及参数配置

本文实验环境如下：CPU 为 Intel Core i7 8700K，GPU 为 GeForce GTX 1080ti，内存大小为 DDR4 32GB，开发环境为 linux 64 位系统，pytorch 1.5.0。

实验参数设置：词嵌入使用 BERT 预训练模型获得，词嵌入向量维度 dbert 为 768；特征提取层，LSTM 隐藏单元数为 256，GAT 卷积核大小为 5，优化函数选择为 Adam，实验学习率初始化为 0.002。实验引入 Dropout_rate = 0.3，提高模型泛化和减少过拟合风险。对文档进行了最大值为 200 个 epochs 的训练与学习，如果到了连续 10 个 epoch 训练中，验证损失都没有减少的话，则终止训练。

4.2. 实验数据集

本文实验的数据来自于上海大学语义智能研究院根据国务院颁布的《国家突发公共事件总体应急预案》的分类体系所构建的中文突发事件预料库(CEC)，该数据集共包含 332 篇突发事件语料，包括车祸、地震、火灾、食物中毒、恐怖袭击五个突发事件类别。

4.3. 评价指标

本文使用准确率(Precision)、召回率(Recall)、F1 值(F-Measure)来评价事件论元识别的效果，Precision、

Recall、F1 定义如(9)~(11)所示：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{Reall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$\text{F1} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{11}$$

4.4. 对比实验分析

为了综合分析模型的特性以及表现，选用一下模型进行对比实验：

DMCNN：基于动态多池网络进行事件论元识别。

JRNN：基于双向 RNN 模型进行事件论元识别。

BiLSTM + CNN：使用 BiLSTM 作为上下文信息提取，并用 CNN 进行特征聚合，最终实现事件论元识别。

Doc2EDGA：预先定义事件抽取框架，使用三个 Transformer 模型进行实体编码，最终通过有向无环图的方式进行事件论元识别。

本文提出的 BiLSTM-GAT 模型与上述基线模型在 CEC 数据集上进行对比实验，其结果如表 1 所示：

Table 1. Model performance comparison
表 1. 模型性能对比

Model	Precision	Recall	F-Measure
DMCNN	0.672	0.648	0659
JRNN	0.682	0.657	0.669
BiLSTM + CNN	0.681	0.691	0.685
Doc2EDGA	0.711	0.690	0.712
BiLSTM-GAT (本文方法)	0.732	0.720	0.720

实验结果表明，本文提出的 BiLSTM-GAT 模型在 Precision，Recall，F-Measure 三个方面都有所提升。这说明 GAT 能够捕获事件句中事件论元之间的语法依存关系，进而提升事件论元识别的效果。

5. 总结

本文针对事件论元识别提出了 BiLSTM-GAT 模型。模型首先使用 BERT 作为词嵌入层，生成文本的语义向量表示；然后通过 BiLSTM 层获得文本中的前向语义信息和后向语义信息；之后通过 GAT 层，获得事件句中论元之间的相互依存关系；最后通过 CRF 层获得论元的序列标注。经过与多个模型进行对比实验，说明本文的模型能够有效地提升事件论元识别的性能。

参考文献

[1] Doddington, G., Mitchell, A., Przybocki, M., *et al.* (2004) The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. *Proceedings of the* 2004 *International Conference on Language Resources and Evaluation*, Lisbon, September 2004, 837-840.

[2] 高李政, 周刚, 罗军勇, 兰明敬. 元事件抽取研究综述[J]. 计算机科学, 2019, 46(8): 9-15.

[3] Ahn, D. (2006) The Stages of Event Extraction. *Proceddings of the Workshop on Annotating and Reasoning about Time and Events*, Sydney, July 2006, 1-8.

[4] McClosky, D., Surdeanu, M. and Manning, C.D. (2011) Event Extraction as Dependency Parsing. *Proceedings of the* 49*th Annual Meeting of the Association for Computational Linguistics*: *Human Language Technologies*, Portland, 1 June 2011, 1626-1635.

[5] Li, P.F., Zhu, Q.M. and Zhou. G.D. (2013) Joint Modeling of Argument Identification and Role Determination in Chinese Event Extraction with Discourse-Level Information. *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, Suzhou, 3 August 2013, 2120-2126.

[6] 秦彦霞, 张民, 郑德权. 神经网络事件抽取技术综述[J]. 智能计算机与应用, 2018, 8(3): 1-5+10.

[7] Chen, Y., Xu, L., Liu, K., *et al*. (2015) Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks. *Proceedings of the* 53*rd Annual Meeting of the Association for Computational Linguistics and the* 7*th International Joint Conference on Natural Language Processing*, Beijing, July 2015, 167-176. https://doi.org/10.3115/v1/P15-1017

[8] Sha, L., Qian, F., Chang, B.B. and Sui, Z.F. (2018) Jointly Extraction Event Triggers and Arguments by Dependency-Bridge RNN and Tensor-Based Argument Interaction. https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16222/16157

[9] Lin, H.Y., Lu, Y.J., Han, X.P., Sun, L. (2016) A Convolution Bilstm Neural Network Model for Chinese Event Extraction. https://eprints.lancs.ac.uk/id/eprint/83783/1/160.pdf

[10] Yang, S., Feng, D.W., Qiao, L.B., Kan, Z.G. and Li, D.S. (2019) Exploring Pre-Trained Language Models for Event Extraction and Geenration. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, July 2019, 5284-5294. https://doi.org/10.18653/v1/P19-1522

[11] Zheng, S., Cao, W., Xu, W. and Bian, J. (2019) Doc2EDAG: An End-to-End Document-Level Framework for Chinese Financial Event Extraction. *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), Hong Kong, November 2019, 337-346. https://doi.org/10.18653/v1/D19-1032