



小型微型计算机系统

Journal of Chinese Computer Systems

ISSN 1000-1220, CN 21-1106/TP

《小型微型计算机系统》网络首发论文

题目：一种基于异构图网络的多模态实体识别方法
作者：李代祎, 张笑文, 严丽
收稿日期：2023-04-17
网络首发日期：2023-07-11
引用格式：李代祎, 张笑文, 严丽. 一种基于异构图网络的多模态实体识别方法[J/OL]. 小型微型计算机系统.
<https://kns.cnki.net/kcms2/detail/21.1106.TP.20230711.1048.002.html>



网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

一种基于异构图网络的多模态实体识别方法

李代祎^{1,2}, 张笑文¹, 严丽¹

¹ (南京航空航天大学 计算机科学与技术学院, 南京 211106)

² (郑州轻工业大学 计算机与通信工程学院, 郑州 450000)

E-mail: lidaiyi@nuaa.edu.cn

摘要: 基于图像信息的辅助, 提高从非结构化文本中识别命名实体的准确率, 可以有效缓解社交媒体场景中因短文本语义信息不全而产生歧义, 图片多却不能发挥作用的问题。尽管现有的研究通常采用跨模态注意力机制合并文本和图像的语义表示, 但是大多不能建立一个一致的表示来融合两种模态之间的语义信息, 且图像中的冗余信息往往会影响多模态实体识别 (Multimodal Name Entity Recognition, MNER) 的性能。为了解决这些问题, 本文提出了一种基于异构图模型的 MNER 方法, 可以有效利用文本和图像之间的交互信息。具体地, 首先, 构建了一个基于 BERT-BiLSTM-CRF 的实体识别模型, 识别出文本中可能存在的实体; 其次, 以文本中可能存在的实体作为两个模态之间的桥梁, 设计了一个由 Token、实体和视觉对象组成的异构图网络, 并定义了两种边来表示相互间的语义关系; 最后, 基于文本和图像组成的异构图, 设计了一种多模态融合模型 (MHGT), 从而减轻了图像噪声的负面影响。在两个通用的 MNER 数据集上的实验结果表明, 本文提出的多模态实体识别方法在 Twitter2015 和 Twitter2017 上分别获得了 75.26% 和 86.51% 的 F1 值, 优于基线模型的性能。

关键词: 多模态实体识别; 注意力机制; 异构图模型; BERT; 条件随机场

中图分类号: TP391

文献标识码:

A

A Multimodal Name Entity Recognition Method Based on Heterogeneous Graph Network

LI Dai-yi^{1,2}, ZHANG Xiao-wen¹, YAN Li¹

¹ (College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

² (College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450000, China)

Abstract: With the aid of image information, improving the accuracy of identifying entities from unstructured text can effectively alleviate the problem of ambiguity caused by incomplete semantic information in short text in social media scenarios, and solve the problem of too many images but not functioning. Although the existing research often used cross-modal attention mechanism to merge the semantic representations of text and images, most of them cannot establish a consistent representation to fuse the semantic information between the two modes, and the redundant information in images often affects the performance of multimodal name entity recognition (MNER). To address these problems, this paper proposes a MNER method based on heterogeneous graph network, which can effectively utilize the interactive information between text and images. Specifically, firstly, an entity recognition model (BERT-BiLSTM-CRF) is constructed to identify the possible entities in the text; Secondly, a heterogeneous graph network consisting of Tokens, entities and visual objects is designed using the possible entities in the text as a bridge between the two modalities, and two edges are defined to represent the semantic relationships between them; Finally, a multimodal fusion model (MHGT) was designed based on heterogeneous graph composed of text and images, thereby reducing the negative impact of image noise. Experimental results on two publicly MNER datasets show that the proposed MNER method achieved 75.26% F1 on Twitter2015 and 86.51% F1 on Twitter2017, respectively, which are superior to the performance of the baseline models.

Key words: Multimodal entity recognition; Attention mechanism; Heterogeneous graph network; BERT; Conditional random field

1 引言

多模态命名实体识别 (Multimodal Name Entity Recognition, MNER) 是信息抽取 (Information Extraction, IE) 领域一项重要任务, 旨在以图像信息作为辅助, 通过补充文本的语义信息来进行消歧, 进而提高 NER 模型的性能。与传统单模态的 NER 方法^[1, 2]不同, 目前 MNER 任务主要面临两个挑战: 1) 如何有效获取与文本中实体相关的视觉信息 (图像中对象); 2) 如何减少图像中冗余信息对 NER 模型的影响。如图 1 所示, 基于文本的语义信息, 文本中的 “Oscars” 可能是一个人名或者是一个电影奖。然而, 图像中的奖杯信息可以补充文本语义的不足, 从而确定文本中的 “Oscars” 是一个电影奖。因此, 针对多模态实体识别任务, 从图像中捕获与实体相关的有效视觉信息是必不可少且具有挑战性的。同时, 如果直接将图像信息和文本信息融合, 图像中的冗余信息可能干扰文本中非实体的标记。



Text: Attenborough and Ben Kingsley with their Oscars.

图 1 Twitter2015 中一个文本和图像对的实例

Fig.1 An example of text-image pair existed in Twitter2015

近年来, 为了有效提高 MNER 模型的性能, 许多研究者通过改进跨模态注意力机制^[3], 有效融合文本和图像信息, 进而获取了进展^[4-6]。尽管这些方法可以直观地通过图像中的视觉特征提高文本中实体识别的真确性, 但是这些方法却忽略了实体语义的完整性, 在符号层面进行跨模态交互, 从而使得文本中非实体标注容易受到图像中冗余信息的干扰。同时, 也有一些研究者通过捕获更好的视觉特征, 从而提高 MNER 任务的准确率^[7-9]。尽管这些方法通过获取有效的视觉特征, 减少了图像中冗余信息的干扰, 但是如何建立一个统一的表示来融合两个模态间的语义信息, 还是一个巨大的挑战。因此, 获取实体关联的有效视觉特征和有效融合文本和图像对象的语义信息, 是 MNER 模型性能提升的关键所在。

综上所述, 基于目前 MNER 任务所面临的问题和挑战。本文提出了一种基于异构图模型的 MNER 方法, 可以有效利用文本和图像之间的交互信息, 从而提高 MNER 模型的性能。该方法主要以实体作为文本和图像两个模态之间的桥梁, 以便从图片对象特征中挖掘与实体相关的视觉特征, 主

要的优势有: 1) 将图像中对象特征与实体表征进行交互, 而不是直接与文本中所有 Token 进行交互, 这可以有效减少图像冗余信息对文本中非实体标记的干扰; 2) 文本中的实体表征具有完全的实体语义, 能够有效获取与实体相关的语义信息。具体地, 首先, 通过构建实体识别模型获取文本中可能实体的表征; 其次, 构造一个有 Token、实体和视觉对象节点组成的异构图, 并定义了两种边来表示相互间的语义关系; 然后, 通过设计的多模态融合模型 (MHGT), 获取与实体相关的视觉特征, 并将其与实体节点想连接的 Token 节点进行融合。最后, 采用 CRF 获取文本中的实体标注序列。本文的贡献主要有:

(1) 为了获取文本中可能实体的表征, 构建了一个实体识别模型 BERT-BiLSTM-CRF。获取的实体表征可以作为两个模态间的桥梁, 方便捕获与实体相关的视觉信息。

(2) 基于 Token、可能实体和视觉对象组成的异构图, 设计了一个多模态融合模型 (MHGT), 可以同时对文本和图像进行表征, 有效缓解了图像噪声的影响。

(3) 在两个广泛使用的多模态数据集 Twitter2015 和 Twitter2015 上进行了实验, 分别获得了 75.26% 和 86.51% 的 F1 值, 验证了本文提出的 MNER 方法的有效性。

2 相关工作

多模态命名实体识别 (MNER) 已经成为命名实体识别 (NER)^[10, 11]领域的一个重要研究方向。因为在许多情况下, 图像与文本是一起出现的 (例如, 社交媒体和购物评论等)。然而, 如何利用视觉信息提高实体识别的准确性, 是当前一个巨大的挑战。因此, 本节主要概述了当前 MNER 任务的相关研究, 并针对当前 MNER 面临的一些挑战和问题提出了相应的解决方法。

一些研究表明, 图像中的视觉信息有助于加强对文本中内容的理解, 从而可以缓解因短文本中语义不足而导致歧义的问题。例如, Moon 等人^[12]最先开展了关于 MNER 的研究, 提出了一个基于 LSTM-CNN-CRF 的 MNER 模型, 通过通用模态注意力模块将文本和图像信息进行融合; 为了减少文本模态和图像模态交互建模过程中图像冗余信息带来的影响, Zhang 等人^[4]提出了一种基于自适应的注意力模型从图像中提取出与文本最相似的视觉特征。Lu 等人^[5]设计一种视觉注意力机制, 可以动态地对文本特征和视觉特征进行融合; 为了使得模态之间的表示更加一致, Yu 等人^[6]设计了一种多模态交互模块, 可以有效捕获图像感知的单词表征和单词感知的图像表征。Xu 等人^[13]通过对比学习^[14]构建了一个通用的文本和图像匹配与对齐框架 (MAF), 可以确定保留视觉信息的比例, 且使得两种模态的表示更加一致。Wu 等人^[7]采用对象标签对文本和图像之间的相互作用进行建模, 并引入了密集的共同注意力机制进行细粒度交互。

尽管上述的方法推动了 MNER 研究的发展, 获得了显著的成果, 但是在一次操作中融合不同模态中的信息, 使得文本和图像的表达不一致, 且图像冗余会直接影响 MNER 模型性能。因此, 一些研究者使用图神经网络 (Graph Neural Network, GNN) [15] 来实现文本和图像的信息融合。例如, Zhang 等人 [9] 在 Yin 等人 [16] 研究的基础上, 提出了一种用于 MNER 的多模态融合方法, 通过图形结构的向量化获取文本和图像的统一表示。然而, 该方法是将图像特征完全和文本特征融合, 却忽略了图像冗余信息的影响。Liang 等人 [17] 设计了一个多模态融合机制, 通过文本和图像对象之间的相似度计算, 完成多模态讽刺检测; Sun 等人 [18] 将文本和图像关系传播的方式引入到预训练 BERT 模型中, 并设计了一种多任务算法来训练和验证关系传播对 MNER 任务的影响。然而, 预训练模型在图像对象检测和词相似性计算方法偏差会直接影响 MNER 的性能。

除此之外, 一些研究旨在通过增强图像的代表能力来提高 MNER 的性能。例如, Wang 等人 [19] 构建了一种新 MNER 模型 (CAT-MNER), 通过整合图像的多粒度信息来加强图像的代表, 从而提高了模型的准确率; Chen 等人 [20] 提出了一种新的分层视觉融合方法, 有效增强的视觉表征, 从而实现了更好的性能。然而, 这些模型仅仅关注的是图像的处理过程, 却忽略了文本和图像之间的交互。

综上所述, 上述的大多数方法将图像的全局信息与文本信息进行交互, 忽略了图形噪声的影响。同时, 考虑到 HGT 不仅可以有效学习异构图中不同节点类型之间的交互信息, 而且可以模仿 Transformer 的结构提取深层次的特征, 进而增强异构图的表示能力。因此, 为了减少实体的 Token 受到图像冗余特征的干扰, 且有效对两个模态进行融合, 本文将文本中可能的实体作为文本和图像之间的桥梁, 通过设计的多模态融合模型 (MHGT), 获取与实体相关的视觉特征, 并将其与实体节点想连接的 Token 节点进行融合, 可以有效减少视觉特征噪声的影响。

3 基于异构图网络的 MNER 模型

在本节中, 首先给出 MNER 任务的定义, 其次详细描述针对 MNER 任务提出的解决模型。

任务定义: 假设给定一个句子 S 以及其关联的图像 I , MNER 的目标是对句子中涉及到的命名实体进行识别并分类到预先定义的实体类别中。实体类别有: 人物 (PER)、组织 (ORG)、位置 (LOC) 和其它 (OTHER) 等。借鉴前人研究工作, 本文将 MNER 任务转化为一个序列标注任务。具体地, 假设 $S = \{x_1, x_2, \dots, x_n\}$ 表示输入句子的单词序列, 并且 $Y = (y_1, y_2, \dots, y_n)$ 是输入序列 S 对应的标注序列, 其中 $y_i \in Y$ 是采用 BIO 标记模式的预定义标签。

3.1 MNER 模型框架

本文提出了一个基于异构图模型的 MNER 方法 (框架

如图 2 所示), 主要分为头实体特征提取模块、图片对象特征编码模块、异构图交互模块和 CRF 标注模块。具体流程, 首先, 该模型通过设计的 BERT-BiLSTM 模型获取包含深层语义信息的单词特征, 充分考虑到了句子的全局信息和局部信息。其次, 基于由 Token、实体和视觉对象节点组成的异构图, 设计了一个多模态融合模型 (MHGT), 可以有效获取与实体相关的视觉特征。最后, 基于视觉特征感知的文本表征, 采用 CRF 获取文本中的实体标注序列。

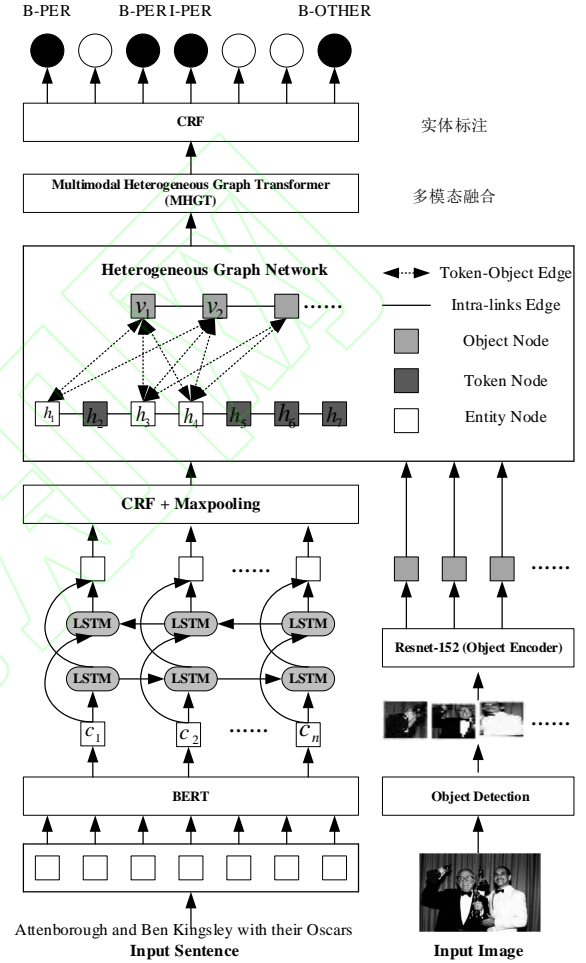


图 2 基于异构图网络的 MNER 框架图

Fig.2 The MNER framework based on MHGT

3.2 基于 BERT-BiLSTM-CRF 的实体表征

基于 BERT 的词嵌入: 假设给定句子表示为 $S = \{x_1, x_2, \dots, x_n\}$, 其中 x_i 表示句子中第 i 个 Token 和 n 表示句子最大的长度。本文采用 Devlin 等人 [21] 提出的 BERT 模型 (核心部件 Transformer 结构如图 3 所示) 去捕获包含深层语义信息的 Token 嵌入序列 $C = \{c_1, c_2, \dots, c_n\}$, 其中 c_i 表示句子中第 i 个 Token 嵌入和 n 表示 Token 的数量。具体的计算如下所示:

$$C = \{c_1, c_2, \dots, c_n\} = \text{BERT}(\{x_1, x_2, \dots, x_n\}) \quad (1)$$

其中, $C \in \mathbb{R}^{n \times d}$ 是输入句子的嵌入表示, d 为隐藏层输出维度, n 为句子的长度。

基于 BiLSTM-ATT 的特征提取: 为了充分考虑句子中

全局信息，将句子嵌入 C 作为 BiLSTM 的输入进行训练。具体地，将两个向相反方向传播的 LSTM 隐藏层的输出特征进行连接，获取具有全局语义信息的文本表示。针对每一个 Token 嵌入 c_i ，前向传播的 LSTM 将学习从第 1 到第 i 个单词的全局信息，标记为 \overrightarrow{LSTM}_i ；以此类推，后向传播的 LSTM 将考学习第 i 个到底 1 个单词的全局信息，标记为 \overleftarrow{LSTM}_i 。经过 BiLSTM 模型训练学习后，句子中第 i 个单词的编码表示如下所示：

$$h_i = [\overrightarrow{LSTM}_i; \overleftarrow{LSTM}_i] \in \mathbb{R}^{n \times 2d}, i = 0, 1, \dots, n \quad (2)$$

$$H = \{h_1, h_2, \dots, h_n\} \in \mathbb{R}^{n \times 2d} \quad (3)$$

其中， $h_i \in \mathbb{R}^{n \times 2d}$ 表示输入句子中每个 Token 的特征表示， d 为隐藏层输出维度， n 为句子的长度。同时，为了捕获句子中关键词的语义信息，通过注意力机制对输入句子中的不同词自动分配不同的权重信息。

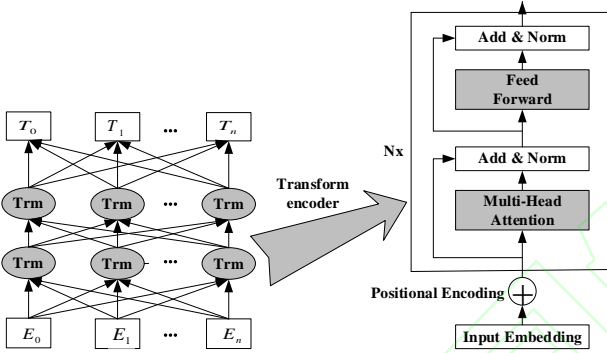


图 3 Transformer 编码器结构

Fig.3 Transformer encoder structure

注意力层：例如，假设输入句子为“*Attenborough and Ben Kingsley with their Oscars.*”，其中“*Attenborough*”和“*Ben Kingsley*”表示两个不同的实体，显然实体之间有关系，同时实体中不同词之间也关系紧密，而实体与非实体词之间距离较远，关系较弱。因此，**本文在 BiLSTM 中嵌入注意力机制^[22]**，有效捕获句子中那关键词的语义信息。注意力机制的流程框架如 4 所示，具体流程如下所述：

首先，句子中词与词之间的注意力权重通过两者之间的相似程度确定，而相似度则通过余弦定理获取。假设两个词之间的注意力权重表示为 $\beta_{i,t}$ ，详细的计算如下：

$$\beta_{i,t} = \frac{\exp(\delta_{i,t})}{\sum_{k=1}^n \exp(\delta_{t,k})} \quad (4)$$

$$\delta_{i,t} = \delta(x_i, x_t) = \frac{w(x_i, x_t)}{|x_i||x_t|} \quad (5)$$

其中， $\delta()$ 表示余弦函数，其 $\delta(x_i, x_t)$ 的计算结果越大，表明单词 x_i 和目标单词 x_t 的相似性越高。 w 表示加权系数。

其次，基于注意力权重 $\beta_{i,t}$ ，对 BiLSTM 隐藏层的输出向量 h_i 进行加权并求和，获取一个蕴含上下文信息的特征向量 Ω_t 。

$$\Omega_t = \sum_{i=1}^n \beta_{i,t} h_i \quad (6)$$

最后，将包含上下文信息的特征向量 Ω_t 与 BiLSTM 隐藏层的输出 h_t 进行连接，连接后的向量表示为 $[\Omega_t; h_t]$ 。然后

采用激活函数 $\tanh()$ 对 $[\Omega_t; h_t]$ 进行非线性映射，获得目标单词 h_t 的注意力向量 \hat{h}_t 。具体计算如下所示：

$$\hat{h}_t = \tanh(W[\Omega_t; h_t]) \quad (7)$$

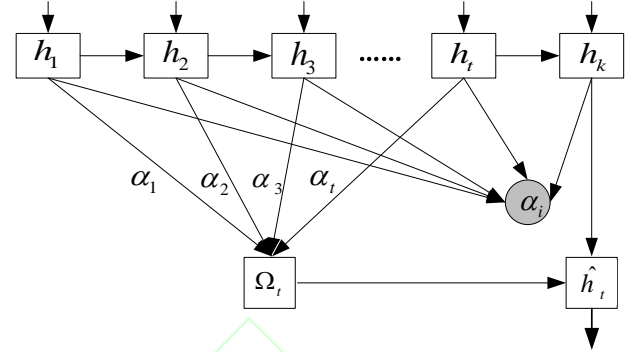


图 4 注意力机制的计算框架

Fig.4 The computational framework of attention mechanism

为了能够和视觉特征进行有效融合，需要将句子中 Token 特征表示的维度转化为多模态空间统一维度，具体的计算如下所示：

$$\bar{H} = W_h \hat{H}^T + b_h \quad (8)$$

$$\bar{H} = \{\bar{h}_1, \bar{h}_2, \dots, \bar{h}_n\} \in \mathbb{R}^{n \times d_m} \quad (9)$$

其中， W_h 和 b_h 分别表示线性变换的权重矩阵和偏置向量； $\hat{H} = \{\hat{h}_1, \hat{h}_2, \dots, \hat{h}_n\}$ 表示经过注意力机制计算后输出的句子特征； d_m 表示多模态空间维度。

基于 CRF-MaxPooling 的实体表示：本文采用 CRF 方法识别输入文本句子中可能存在的实体 $\{(s_i, e_i)\}_{i=1}^{|E|}$ ，其中 $|E|$ 表示识别出的实体数量， s_i 和 e_i 分别表示第 i 个实体的开始和结尾标记。实体标记的损失函数 L_{loss} 计算如下：

$$L_{loss} = -\sum_i \log P(z|S) \quad (10)$$

$$P(z|S) = \frac{\prod_{i=1}^{|n|} \varphi_i(z_{i-1}, z_i, S)}{\sum_{z' \in Z} \prod_{i=1}^{|n|} \varphi_i(z'_{i-1}, z'_i, S)} \quad (11)$$

其中， $\varphi_i(z_{i-1}, z_i, S)$ 和 $\varphi_i(z'_{i-1}, z'_i, S)$ 是 CRF 中的势函数（Potential functions）。为了获取完整的实体表示，通过最大池化操作获得实体中每个 Token 的特征表示，具体表示如下所示：

$$E_n = \text{Max}(\{\bar{h}_i\}_{i=s_n}^{e_n}) \quad (12)$$

其中， E_n 表示实体表征， $\text{Max}()$ 表示最大池化计算， s_n 和 e_n 分别表示第 n 个实体的开始和结尾标记。

3.3 基于残差网络的视觉特征提取

考虑到图片对象具有与实体相似的语义信息，为了减少图片冗余信息对 MNER 的影响，**本文选择图片中对象的表示作为图片的视觉特征**。视觉特征的具体提取过程如下所述：

首先，采用 DERT 模型^[23]完成图片中的对象检测。假设输入图像表示为 I ，则图片中对象 O 表示如下所示：

$$O = [o_1, o_2, \dots, o_m] = \text{DERT}(I) \quad (13)$$

其中，图像中检测出的对象个数用 m 表示， o_i 表示图像

中第 i 个视觉对象，用 2048 维向量表示。

其次，将图像表示 I 和对象表示 O 进行链接，然后将其输入到一个 152 层的 ResNet 中学习每个图像表示，并将 ResNet 中最后一个卷积层的输出 $V = \{v_i\}_{i=1}^{O+1}$ 作为视觉特征，具体的表示如下所示：

$$V = \text{ResNet}([I; O]) \in \mathbb{R}^{(m+1) \times 2048} \quad (14)$$

为了使得视觉特征与多模态空间的维度一致，通过 ReLU 激活函数将视觉特征投影到多模态空间，具体计算如下所示：

$$\bar{V} = \text{ReLU}(W_v V^T + b_v) \in \mathbb{R}^{(m+1) \times d_m} \quad (15)$$

3.4 多模态异构图交互网络 (MHGT)

本文设计了一个 MHGT 模块（如图 5 所示），可以通过实体和图片对象的交互作用获取实体感知的视觉信息，并将其融合到相关实体的 Token 表示中。与一次向量化视觉节点和文本节点的研究不同，MHGT 模块可以直接对构建的异构图进行向量化，可以同时获取视觉节点和文本节点的特征向量化表示。

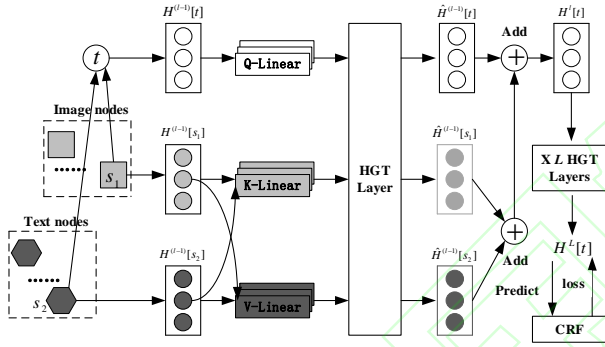


图 5 MHGT 的计算框架

Fig.5 The calculation process framework of MHGT

异构图构建：以文本节点和视觉对象节点为基础，构建了一个异构图 G 。该多模态异构图主要包含三种类型节点：Token 节点 $N_{T_i} = \bar{h}_i$ 、实体节点 $N_{E_i} = E_i$ 和视觉节点 $N_{V_i} = \bar{v}_i$ ；异构图中的边有两种类型：实体与对象之间的边，以便有效捕获实体感知的视觉信息，减少图像噪声的影响；模态内节点互相连接的边，相同模态中任意两个节点可以通过内部链路连接，可以将全局信息引入到异构图中。异构图的形式化表示如下所示：

$$G = (V, \Delta(V), f(V), E, \Delta(E), f(E)) \quad (16)$$

其中， V 表示节点的集合； E 表示边的集合； $\Delta(V)$ 和 $\Delta(E)$ 分别表示节点和边的类型； $f(V)$ 和 $f(E)$ 分别表示节点和边的特征。

多模态 HGT 构建：为了能够在保持异构图中节点和边类型相关的表示，同时对不同类型的节点和边能够同时进行建模，捕获异构图的属性。本文采用 HGT^[24]对文本和图像构建的异构图进行建模，获取对象视觉信息感知的文本表示，进而提高 MNER 任务的准确性。MHGT 的核心思想是通过聚集来自源点的信息获取每个节点的表示，本文采用 HGT 学习每个样本图的上文表示，其主要的优势有：1) 通过 HGT 可以学习局部图的表示，每个局部图包含目标节点 t ，

以及与其相邻节点 $s_i (i = 1, 2, \dots, n)$ ；2) 在 HGT 学习样本图的过程中，可以使得不同类型的节点和边保持各自的空间表示，同时不同类型的节点还可以通过消息专递进行交互学习，进而保证语义信息和视觉信息的完整性。

针对原始异构图的采样图，HGT 可以对所有连接的节点对进行建模学习，其中目标节点 t 通过边 e 与源节点 s 相连接。通过 MHGT 获取目标节点 t 的上文表示。具体计算如下所述：

首先，基于 HGT 计算目标节点 t 和源节点 s 之间的注意力。具体地，针对每一个边 $e = (s, t)$ 计算 k 头注意力，计算如下所示：

$$ATT_{(s,e,t)} = \text{softmax}_{V \in N(t)} (\prod_{i \in [1,k]} ATT_head^i(s, e, t)) \quad (17)$$

$$ATT_head^i(s, e, t) = (K^i(s) W_{\Delta(e)}^{ATT} Q^i(t)^T) \cdot \frac{\mu}{\sqrt{d}} \quad (18)$$

$$k \leftarrow K^i[s] = K - \text{Linear}_{\Delta(s)}^i(H^{l-1}[s]) \quad (19)$$

$$q \leftarrow Q^i[t] = Q - \text{Linear}_{\Delta(t)}^i(H^{l-1}[t]) \quad (20)$$

$$v \leftarrow V^i[s] = V - \text{Linear}_{\Delta(s)}^i(H^{l-1}[s]) \quad (21)$$

其中， $ATT_head^i(s, e, t)$ 表示第 i 个头注意力； $W_{\Delta(e)}^{ATT}$ 表示权重矩阵； μ 表示先验张量，用于对注意力的自适应缩放； $K^i[s]$ 、 $Q^i[t]$ 和 $V^i[s]$ 分别表示第 i 个头的查询向量、键向量和值向量。此外， H^{l-1} 表示 HGT 中第 $l-1$ 层的输出节点特征； $K - \text{Linear}_{\Delta(s)}^i$ 、 $Q - \text{Linear}_{\Delta(t)}^i$ 和 $V - \text{Linear}_{\Delta(s)}^i$ 是将不同类型的节点特征映射到同一特征空间的线性变换。

其次，计算节点间互注意力的同时，需要将信息从源节点传递到目标节点。针对每一个边 $e = (s, t)$ ，其多头消息的计算如下所示：

$$MSG_{(s,e,t)} = \prod_{i \in [1,k]} MSG_head^i(s, e, t) \quad (22)$$

$$MSG_head^i(s, e, t) = \text{Linear}_{\Delta(s)}^i(H^{l-1}[s]) W_{\Delta(e)}^{MSG} \quad (23)$$

其中， $MSG_head^i(s, e, t)$ 表示第 i 个消息头；线性变换 $M - \text{Linear}_{\Delta(s)}^i$ 的目的是将源节点 s 的特征投影到第 i 个消息头向量中； $W_{\Delta(e)}^{MSG}$ 表示权重矩阵。

最后，公式 (17) 中的 Softmax 函数已经使得每个目标节点的 t 的注意力向量的总和为 1。因此，本文将简单地使用注意力向量作为权重，对来自源节点的消息进行融合，获取 HGT 中第 l 层的输出，具体计算如下：

$$H^l[t] = \bigoplus_{V \in N(t)} (ATT_{(s,e,t)} \cdot MSG_{(s,e,t)}) \quad (24)$$

其中， \bigoplus 表示加和的操作； $H^l[t]$ 表示将不同源节点 s 信息聚合到目标节点 t 后的向量表示。

3.5 CRF 实体标注

本文采用条件随机场 (Conditional Random Fields, CRF)^[25]对输入序列进行标记，既可以充分考虑相邻标签之间的相关性又可以并对整个标签序列进行评分。具体地，假设 $H^l = \{h_1^l, h_2^l, \dots, h_n^l\}$ 表示输入句子的表示， $y = \{y_1, y_2, \dots, y_n\}$ 表示对应的标注序列，其中 $y_i \in Y$ 表示第 i 个词对应标签的编码表示。具体的计算过程如下所述：

$$P(y|X) = \frac{\exp(\text{score}(H^l, y))}{\sum_{\tilde{y} \in Y} \exp(\text{score}(H^l, \tilde{y}))} \quad (25)$$

$$\text{score}(H^l, y) = \sum_{i=0}^n T_{(y_i, y_{i+1})} + \sum_{i=0}^n E_{(h_i^l, y_i)} \quad (26)$$

$$E_{(h_i^l, y_i)} = W_f^{y_i} \cdot h_i^l \quad (27)$$

其中, Y 表示针对输入句子对应的所有可能的标签序列集合; $T_{(y_i, y_{i+1})}$ 表示从标签 y_i 转移到标签 y_{i+1} 的转移分数; $E_{(h_i^l, y_i)}$ 表示第 i 个单词对应标签 y_i 的得分。

在训练过程中, 使用最大似然估计函数对进行标记预测。

$$L(P(y|X)) = \sum_i \log(P(y|X)) \quad (28)$$

在预测过程中, 使用维特比算法(Viterbi)^[26]选出所有预测标注序列中得分最高的标注序列, 并将其作为最终的实体标签序列的标注结果, 具体的计算如下所示:

$$\bar{y} = \underset{y \in Y}{\operatorname{argmax}} (P(y|X)) \quad (29)$$

4 实验

4.1 实验数据与评估指标

实验数据集: 实验中, 本文采用两种广泛使用的 MNER 数据集: Twitter2015^[4]和 Twitter2017^[5]。两个数据集中每一条数据都是由图像和对应文本构成, 文本内容可能不在图像中, 且文本中可能没有实体或包含多个实体。这两个数据集中实体的类型主要有四类: 人物(PER)、组织(ORG)、位置(LOC)和其它(OTHER)。本文采用 Yu 等人^[6]提供的预处理方法对数据进行预处理, 两种数据集的具体统计分析如表 1 所示:

表 1 两个 MNER 数据集分析

Table 1 The statistics summary of two MNER datasets

Type	Twitter2015			Twitter2017		
	Train	Dev	Test	Train	Dev	Test
PER	2217	552	1816	2943	626	621
LOC	2091	522	1697	731	173	178
ORG	928	247	839	1674	357	395
OTHER	940	225	726	701	150	157
Total	6176	1546	5078	6049	1324	1351
Tweets	4000	1000	3257	3373	723	723

其中, 两个数据集中数据主要包括 2014 年到 2015 年和 2016 年到 2017 年两年间用户在 Twitter 上发布的文本和图像数据。

评估指标: 本文采用通用的分类评估指标对构建的 MNER 模型进行性能评估, 主要有精确率(P)、召回率(R)和 $F1$ 评分。

$$P = \frac{TP}{TP+FP} \quad R = \frac{TP}{TP+FN} \quad F_\beta = \frac{(\beta^2+1)PR}{\beta^2P+R} \quad (\beta^2 \in [0, +\infty)) \quad (30)$$

其中, TP 表示正确预测实体的个数; FP 表示错误预测为实体的个数; FN 表示真实实体没被正确预测的个数。

4.2 超参数设置与实验环境

超参数设置: 通过参考模型训练结果和相关文献中的研究, 设置了提出的 MNER 模型的超参数值。目前, 考虑到实验室硬件条件, 本文选择 BERT-Base 模型来捕获输入序列的深层语义特征。同时, 设计嵌入全局注意力机制的 BiLSTM 模型来获取蕴含全局信息的序列表征, 并通过 Adam 优化器^[27]来最小化损失函数。提出的 MNER 模型中超参数的设置如表 2 所示:

表 2 超参数值设置

Table 2 Hyperparametric value setting

参数名	Twitter2015	Twitter2017
最大句子长度	128	128
Batch 大小	16	16
BERT	BERT-Base	BERT-Base
图片对象	≤ 5	≤ 5
ResNet152 维度	2048	2048
多模态空间维度	256	256
HGT 隐藏层维度	128	128
HGT 层数	5	5
HGT 多头个数	2	2
学习率	3e-5	1e-5

实验环境: 在实验中, 本文使用了 Pytorch 框架在 AMD NVidia RTX2080Ti (11GB) GPU 上完成了所有实验。

4.3 实验结果与讨论

实验结果: 为了验证本文提出的 MNER 模型是有效的, 在 Twitter2015 和 Twitter2017 数据集上分别对所提出的模型与两组基线模型进行性能比较。

(1) 基于文本的 NER 基线方法

CNN-BiLSTM-CRF^[28]: 该模型是一种经典的 NER 模型, 可以有效获取输入序列的单词级和字符级的特征信息;

HBiLSTM-CRF^[29]: 该模型是基于 CNN-BiLSTM-CRF 模型的改进。采用 BiLSTM 替换掉了 CNN 层, 以便获取文本更深层次的语义特征;

BERT-CRF^[30]: 该模型是 BERT 的改进, 采用 CRF 替代 Softmax 层对序输入列进行实体标注。

(2) 基于文本和图像的 MNER 基线方法

VG^[5]: 该模型利用视觉注意和门控机制来从整个图像中挖掘深层信息, 指导 HBiLSTM-CRF 模型对文本进行表征学习;

ACN^[4]: 该模型通过设计的自适应共同注意力网络, 导基于 CNN-BiLSTM-CRF 模型对文本进行表征学习, 获取视觉感知的文本表示;

UMT^[6]: 该模型将 Transformer 扩展到多模态版本, 并在 Transformer 中集成了辅助实体跨模态检测模块;

UMGF^[9]: 该模型将图进行矢量化表示, 得到文本和图像的同一维度表示, 然后对两者直接进行融合获取图像感知

的文本表示；

MAF^[13]：该模型提出了一种新的交叉模态匹配模块和一种新的交叉模态对准模块，可以确定保留视觉信息的比例，减少图像噪声的影响。

在 Twitter2015 数据集上，将提出的 MNER 模型与上述基线模型进行性能对比，结果如表 3 和表 4 所示：

表 3 在 Twitter2015 数据集上的模型性能对比

模型	Single Type (F1)			
	PER	LOC	ORG	OTHER
CNN-BiLSTM-CRF	80.86	75.39	47.77	32.61
HBiLSTM-CRF	82.34	76.83	51.59	32.52
BERT-CRF	84.74	80.51	60.27	37.29
VG	82.66	77.21	55.06	35.25
ACN	81.98	78.95	53.07	34.02
UMT	85.24	81.58	63.03	39.45
UMGF	84.26	83.17	62.45	42.42
MAF	84.67	81.18	63.35	41.82
本文	85.16	83.91	64.48	43.65

表 4 在 Twitter2015 数据集上的模型性能对比

模型	Overall		
	P (%)	R (%)	F1 (%)
CNN-BiLSTM-CRF	66.24	68.09	67.15
HBiLSTM-CRF	68.14	61.09	64.42
BERT-CRF	69.22	74.59	71.81
VG	73.96	67.90	70.80
ACN	72.75	68.74	70.69
UMT	71.67	75.23	73.41
UMGF	74.49	75.21	74.85
MAF	71.86	75.10	73.42
本文	75.83	74.69	75.26

在 Twitter2017 数据集上，将提出的 MNER 模型与上述基线模型进行性能对比，结果如表 5 和表 6 所示：

表 5 在 Twitter2017 数据集上的模型性能对比

模型	Single Type (F1)			
	PER	LOC	ORG	OTHER
CNN-BiLSTM-CRF	87.99	77.44	74.02	60.82
HBiLSTM-CRF	87.91	78.57	76.67	59.32
BERT-CRF	90.25	83.05	81.13	62.21
VG	89.34	78.53	79.12	62.21
ACN	89.63	77.46	79.24	62.77
UMT	91.56	84.73	82.24	70.10

UMGF	91.12	85.22	83.13	69.83
MAF	91.51	85.80	85.10	68.79
本文	92.35	86.22	84.18	68.45

表 6 在 Twitter2017 数据集上的模型性能对比

Table 6 Comparison of model performance on the Twitter 2017

模型	Overall		
	P (%)	R (%)	F1 (%)
CNN-BiLSTM-CRF	80.00	78.76	79.37
HBiLSTM-CRF	82.69	78.16	80.37
BERT-CRF	83.32	83.57	83.44
VG	83.41	80.38	81.87
ACN	84.16	80.24	82.15
UMT	85.28	85.34	85.31
UMGF	86.54	84.50	85.51
MAF	86.13	86.38	86.25
本文	87.29	85.75	86.51

结果讨论：本文在两个基准 MNER 数据集上，针对提出的 MNER 模型和基线模型，分别给出了在每种实体类型上的 $F1$ 评分和总体上的精确率 (P)、召回率 (R) 和 $F1$ 评分。具体分析如下所述：

首先，将本文提出的 MNER 模型与一些基于文本的 NER 模型进行比较。如上述 4 个表所示，可以显然发现基于 BERT 的方法表现较好，这表明基于迁移学习的预训练模型不是从头开始训练学习，有助于提升 NER 任务的准确率。此外，BERT-CRF 模型比仅使用 BERT 模型具有更好的性能，表明 CRF 确实可以有效地学习邻域中标签的约束，并联合预测最佳标签序列。

其次，将 MNER 方法（如表 4 和表 6）与基于文本的单模态方法（表 3 和表 5）进行比较，可以明显发现几乎所有的 MNER 模型都优于其相应的基文本的单模态 NER 模型，这表明社交媒体 Twitter 上的图像信息确实有助于文本中命名实体的识别。

最后，将本文提出的 MNER 模型与其它一些 MNER 模型进行比较。如上述 4 个表所示，显然本文提出的 MNER 方法在两个数据集上都达到了最先进的性能，有效证明了本文提出 MNER 模型的有效性。特别是在 TWITTER-2017 数据集上，本文提出的 MNER 模型在总体上的 $F1$ 上优于最先进的模型（MAF），这表明本文提出的 MHGT 有助于模型更好地融合文本和图像的表达。

除此之外，在 Twitter2015 数据集上，尽管 UMT 模型在标签为“PER”实体的识别获得最高的 $F1$ 值（85.24%），但是本文提出的模型在整体性能上明显优于 UMT（如表 4 所示）；在 Twitter2017 数据集上，尽管 UMT 和 MAF 在标签为“ORG”和“OTHER”实体的识别中分别获得最高的 $F1$ 值（85.10%和 70.10），且在多模态实体识别中 MAF 获

得了较高的召回率 P (86.38%),但是在整体性能上本文提出模型是最优的(如表6所示)。导致上述结果的主要原因是本文模型能够更加充分利用有效的视觉信息,减少图像噪声的影响,同时也可能是因为:1)数据集中实体类型的分布不均;2)不同数据集的有偏标记。

4.4 消融实验

为了研究本文提出的 MNER 模型中各个模块的有效性,在两个基准数据集上进行了相关消融实验。该模型的主要模块有实体特征提取模块、图片对象特征编码模块和异构图交互模块,具体分析如下所述:

表 7 不同模型中各个模块的作用

Table 7 The role of each module in our model

模型	Twitter2015			Twitter2017		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
w/o BERT	68.54	69.05	68.79	79.68	74.42	76.96
w/o ResNet	69.88	70.62	70.25	83.05	82.64	82.84
w/o MHGT	73.67	71.55	72.59	85.12	83.85	84.48
本文	75.83	74.69	75.26	87.29	85.75	86.51

如表7所示,实验结果表明本文提出的 MNER 模型中的各个模块都对最终结果做出了重要贡献。1)去除 BERT 模块后,模型性能明显下降,表明了 BERT 有助于模型获取文本深层次的语义特征;2)去除 ResNet152 模块,在没有图像信息辅助的情况下,NER 模型的性能下降,这表明视觉信息确实有助于文本中命名实体的识别。3)去除 MHGT 模块,就是将图像信息和文本信息通过跨模态注意力机制进行融合,而不通过图神经网络对异构图进行建模。NER 模型性能显著下降,这表明同时对不同类型的节点和边进行建模,捕获异构图的属性,可以有效改善 MNER 模型性能。

4.5 实例分析

为了更好地了解相关视觉信息是否有助于文本中命名实体的识别,本文从两个基准数据集中选择了两个具有代表性的测试样本,比较了不同方法的预测结果。



首先,通过观察发现,多模态实体识别方法能够获得较好的结果,往往是因为提供的文本数据不正确或不完整,而图片中的视觉信息却可以提供一些有用的线索。例如,在表8(A)中,显然在没有视觉信息支持的情况下,BERT-CRF 无法识别出两个实体是指音乐会中的两个歌手,但是提供的多模态实体识别方法却可以正确地识别出文本中的实体,并能进行正确的分类。

其次,通过查看统计两个基准数据集上的测试样例,发现大约5%的社交媒体帖子中,图像可能与文本内容不相关,主要原因有:1)这些社交媒体的帖子中包含模糊图像、漫画或带有隐喻的照片;2)图片中的视觉信息和文本中的内容表示了同一事件的不同侧面。在这种情况下,通过观察发现,多模态实体识别的方法通常比 BERT-CRF 模型表现

更差。例如,在表8(B)中,图像与文本内容反映的不是同一件事,但是图像中存在人物头像,提供的多模态实体识别方法都错误地将“Siri”分类为“PER”。因此,图像的噪声对文本实体识别是有直接影响的。

表 8 测试样例上的实体预测

Table 8 Entity prediction on test samples

模型	相关图像的重要性	图像噪声的影响
		
	A. My mom took some awesome photos of imarationale and bastilledan .	B. Ask Siri what 0 divided by 0 is and watch her put you in your place.
BERT-CRF	1-OTHER, 2-ORG (错误)	1-OTHER(正确)
MHGT	1-PER, 2-PER (正确)	1-PER(错误)
本文	1-PER, 2-PER (正确)	1-PER(错误)

如表8所示,与文本相关的图像可以辅助文本中的实体被正确识别,例如,“imarationale”和“bastilleda”被 MNER 模型正确识别为 PER。同时,图像的噪声也会直接影响模型性能,例如,由于图像中人物头像的视觉信息,使得“Siri”被 MNER 模型错误识别为 PER。

5 结论

本文提出了一个基于异构图网络的 MNER 模型,可以有效利用文本和图像之间的交互信息。具体地,为了获取包上下文信息的实体表示,构建了一个基于 BERT-BiLSTM-CRF 的实体识别模型;此外,为了获取与实体相关的视觉信息,以及减少图像噪声的影响,设计了一个由 Token、实体和视觉对象组成的异构图,并通过构建的 MHGT 模型对视觉信息和文本信息进行交互建模。在基准数据集上的实验结果表明,本文提出的 MNER 模型是有效且稳定的。

这项工作未来的几个研究方向。一方面,尽管现有的 MNER 方法已经取得显著成果,但是在图像和文本不匹配的情况下,MNER 模型性能仍然表现不佳。因此,研究动态去除图像中潜在噪声,进一步提高 MNER 模型性能。另一方面,由于现有 MNER 数据集中数据量较少,限制了 MNER 任务的研究,下一步将利用不同社交平台中大量未标记的社交帖子,结合少量标注数据,训练更强大的 MNER 模型。

References:

[1] Li J, Sun A, Han J, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(1): 50-70.

- [2] Li D, Yan L, Yang J, et al. Dependency syntax guided bert-bilstm-gam-crf for chinese ner[J]. Expert Systems with Applications, 2022, 196: 116682.
- [3] Cao Jian-jun, Nie Zi-bo, Zheng Qi-bin, et al. A review of deep learning entity relationship extraction research[J]. Journal of Software, 2023, 11(4): 1-26.
- [4] Zhang Q, Fu J, Liu X, et al. Adaptive co-attention network for named entity recognition in tweets[C]//AAAI Conference on Artificial Intelligence, 2018, 32(1).
- [5] Lu D, Neves L, Carvalho V, et al. Visual attention model for name tagging in multimodal social media[C]//56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018: 1990-1999.
- [6] Yu J, Jiang J, Yang L, et al. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer[C]//Association for Computational Linguistics, 2020: 3342-3352.
- [7] Wu Z, Zheng C, Cai Y, et al. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts[C]//28th ACM International Conference on Multimedia, 2020: 1038-1046.
- [8] Chen S, Aguilar G, Neves L, et al. Can images help recognize entities? A study of the role of images for Multimodal NER[C]//7th Workshop on Noisy User-generated Text (W-NUT 2021), 2021: 87-96.
- [9] Zhang D, Wei S, Li S, et al. Multi-modal graph fusion for named entity recognition with targeted visual guidance[C]//AAAI conference on artificial intelligence, 2021, 35(16): 14347-14355.
- [10] Mai C, Liu J, et al. Pronounce differently, mean differently: a multi-tagging-scheme learning method for Chinese NER integrated with lexicon and phonetic features[J]. Information Processing & Management, 2022, 59(5): 103041.
- [11] He Q, Wu L, et al. Knowledge-graph augmented word representations for named entity recognition[C]//AAAI Conference on Artificial Intelligence, 2020, 34(5): 7919-7926.
- [12] Moon S, Neves L, Carvalho V. Multimodal named entity disambiguation for noisy social media posts[C]//56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018: 2000-2008.
- [13] Xu B, Huang S, et al. MAF: a general matching and alignment framework for multimodal named entity recognition[C]//15th ACM Conference on Web Search and Data Mining, 2022: 1215-1223.
- [14] Bhattacharjee A, Karami M, Liu H. Text transformations in contrastive self-supervised learning: a review[C]//31st International Joint Conference on Artificial Intelligence, IJCAI 2022, International Joint Conferences on Artificial Intelligence, 2022: 5394-5401.
- [15] Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(1): 4-24.
- [16] Yin Y, Meng F, Su J, et al. A novel graph-based multi-modal fusion encoder for neural machine translation[J]. arXiv preprint arXiv:2007.08742, 2020.
- [17] Liang B, Lou C, Li X, et al. Multi-modal sarcasm detection via cross-modal graph convolutional network[C]//60th Annual Meeting of the Association for Computational Linguistics, 2022: 1767-1777.
- [18] Sun L, Wang J, Zhang K, et al. RpBERT: a text-image relation propagation-based BERT model for multimodal NER[C]//AAAI Conference on Artificial Intelligence, 2021, 35(15): 13860-13868.
- [19] Wang X, Ye J, Li Z, et al. Cat-mner: multimodal named entity recognition with knowledge-refined cross-modal attention [C]//IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2022: 1-6.
- [20] Chen X, Zhang N, Li L, et al. Good visual guidance makes a better extractor: hierarchical visual prefix for multimodal entity and relation extraction[J]. arXiv preprint arXiv:2205.03521, 2022.
- [21] Devlin J, Chang M W, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
- [22] Li Dai-yi, Li Zhong-liang, Yan Li. Research on chinese-oriented entity relation joint extraction method[J]. Journal of Chinese Computer Systems, 2022, 43 (12): 2479-2486.
- [23] Carion N, Massa F, et al. End-to-end object detection with transformers[J]. arXiv preprint arXiv:2005.12872, 0 2020.
- [24] Hu Z, Dong Y, Wang K, et al. Heterogeneous graph transformer[C]//Web Conference, 2020: 2704-2710.
- [25] Su S, Qu J, Cao Y, et al. Adversarial training lattice lstm for named entity recognition of rail fault texts[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(11): 21201-21215.
- [26] Forney G D. The viterbi algorithm[J]. Proceedings of the IEEE, 1973, 61(3): 268-278.
- [27] Kingma D P, Ba J. Adam: a method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
- [28] Ma X, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF[C]//54th Annual Meeting of

the Association for Computational Linguistics,2016: 1064-1074.

[29] GLample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[J]. arXiv preprint arXiv:1603.01360, 2016.

[30] Sun Hong, Chen Qiang-yue. Chinese text classification integrating Bert word embedding and attention mechanism [J]. Journal of Chinese Computer Systems,2022,43(1):22-26.

附中文参考文献:

[3] 曹建军,聂子博,郑奇斌,等. 跨模态数据实体分辨研究综述[J].软件学报.2023,11(4):1-26.

[22] 李代祎,李忠良,严 丽. 一种面向中文的实体关系联合抽取方法研究[J].小型微型计算机系统,2022,43(12):2479-2486.

[30] 孙 红,陈强越.融合 BERT 词嵌入和注意力机制的中文文本分类[J].小型微型计算机系统,2022,43(1):22-26.