DOI:10.11918/202201126

融合知识图谱的医疗领域命名实体识别

金志刚[1],何晓勇[1],岳顺民[2,3],熊亚岚[1],罗 嘉[1]

(1. 天津大学 电气自动化与信息工程学院,天津 300072; 2. 国网天津市电力公司,天津 300010;
3. 天津市能源大数据仿真企业重点实验室,天津 300010)

摘 要:为了改善通用预训练模型不适应医疗领域的命名实体识别任务这一不足,提出了一种融合医疗领域知识图谱的神经网络架构,该架构利用弹性位置和遮盖矩阵使预训练模型计算自注意力时避免语义混淆和语义干扰,在微调时使用多任务学习的思想,利用回忆学习的优化算法使预训练模型均衡通用语义表达和目标任务的学习,最终得到更为高效的向量表示并进行标签预测。实验结果表明:本文提出的命名实体识别架构在医疗领域上取得了优于主流预训练模型的效果,在通用领域也有较为良好的效果。该架构避免了重新训练针对某个领域的预训练模型和引入额外的编码结构从而精简了计算代价和模型大小。此外,通过消融实验对比,医疗领域对于知识图谱的依赖程度较通用领域依赖程度更大,这说明在医疗领域中融合知识图谱方法的有效性。通过参数分析,证明本文使用回忆学习的优化算法可以有效控制模型参数的更新,使模型可以保留更多的通用语义信息并得到更符合语义的向量表达。本文也通过实验分析说明了所提方法在实体数量少的种类上具有更优的表现。

关键词:BERT;知识图谱;多任务学习;命名实体识别

中图分类号:TP183 **文献标志码**:A **文章编号**:0367-6234(2023)05-0050-09

Named entity recognition in medical domain combined with knowledge graph

JIN Zhigang[1], HE Xiaoyong[1], YUE Shunmin[2,3], XIONG Yalan[1], LUO Jia[1]

(1. School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China;
2. State Grid Tianjin Electric Power Company, Tianjin 300010, China; 3. Key Laboratory of
Energy Big Data Simulation of Tianjin Enterprise, Tianjin 300010, China)

Abstract: In view of the problem that general pre-trained models are not suitable for named entity recognition tasks in the medical domain, a neural network architecture that integrates knowledge graph in the medical domain was proposed. The elastic position and masking matrix were used to avoid semantic confusion and semantic interference in self-attention calculation of pre-trained model. The idea of multi-task learning in fine-tuning was adopted, and the optimization algorithm of recall learning was employed for pre-trained model to balance between general semantic expression and learning of the target task. Finally, a more efficient vector representation was obtained and label prediction was conducted. Experimental results showed that the proposed architecture achieved better results than the mainstream pre-trained models in the medical domain, and had relatively good results in the general domain. The architecture avoided retraining pre-trained models in particular domain and additional coding structures, which greatly reduced computational cost and model size. In addition, according to the ablation experiments, the medical domain was more dependent on the knowledge graph than the general domain, indicating the effectiveness of integrating the knowledge graph method in the medical domain. Parameter analysis proved that the optimization algorithm which used recall learning could effectively control the update of model parameters, so that the model retained more general semantic information and obtained more semantic vector representation. Besides, the experimental analysis showed that the proposed method had better performance in the category with a small number of entities.

Keywords: bidirectional encoder representation from transformers (BERT); knowledge graph; multi-task learning; named entity recognition

随着医疗系统信息化程度不断提高,国家对医疗系统中的数字化过程越来越重视。电子病历(EMRs)收集了医务人员在医疗过程中使用医疗机构信息系统生成的数字化信息。有研究已经说明电

收稿日期:2022-01-28;**录用日期**:2022-04-02;**网络首发日期**:2022-05-05
网络首发地址:https://kns.cnki.net/kcms/detail/23.1235.T.20220505.0943.026.html
基金项目:国家自然科学基金(71502125)
作者简介:金志刚(1972—),男,教授,博士生导师
通信作者:金志刚,zgjin@tju.edu.cn

子病历不仅可以挖掘和患者相关的医疗知识，还可以用来解释或者推断一些有用的信息[1]。自电子病历产生至今，已经积累了大量电子病历数据，提取电子病历中的结构化数据可以用于构建个性化医疗服务体系和临床决策支持[2]，而抽取电子病历中的信息的基础在于准确识别出电子病历中的相关实体。命名实体识别是自然语言处理中的基础任务之一，也是信息抽取中的重要环节。该任务旨在识别文本中属于预先定义的类别的实体并判断其所属类别，一般分为通用命名实体识别和领域命名实体识别。其中通用命名实体识别通常识别人名、地名和日期等一般实体，领域命名实体识别通常识别本专业领域内涉及的实体类型。例如在医学领域中，药物类型、身体部位、疾病类型和症状类型一般是电子病历中经常出现的实体类型，这些实体类型区别于通用领域的命名实体，对医学领域有着独特的作用。

命名实体识别基于深度神经网络的方法以其高效的非线性函数表达能力和自动捕捉潜在特征的能力在识别效果和构建方法上均相较于传统方法有显著的优势[3]。利用神经网络实现命名实体识别的关键在于表示学习[4]，以 Word2vec[5] 算法为代表的静态词向量表示最初被研究者使用。然而自从 ELMo[6]、BERT[7] 预训练模型提出后，利用堆叠神经网络得到文本的动态词向量再进行有监督学习的微调方法成为了新的自然语言处理范式，BERT 是一种使用注意力机制的深层神经网络，注意力机制模仿人类的行为，在计算机视觉中的行人遮挡[8] 等任务上都有出色的表现。

预训练模型已经在包含命名实体识别在内的诸多自然语言处理任务上获得了出色的效果[6-7,9]。预训练模型采用通用语料进行自监督训练，然而通用语料的语言习惯和文本特征与领域文本之间存在着一定的差异。有研究[10] 统计了其构建的领域词表的重复覆盖比例，不同领域的词表覆盖率最低可达到 12.7%；该研究表示：当自监督训练的源任务领域文本和目标任务领域文本所对应的领域不同时，模型的效果下降非常明显。因此预训练模型在通用语料上的预训练并不能很好地适配各种各样的领域自然语言处理任务。文献[11 – 12] 也表明使用通用语料做自监督训练的预训练模型不能很好地适应医学相关领域的文本任务。一种直接的解决方法是构建领域预训练模型，文献[11] 利用医学文本训练生物医学领域的 ELMo 预训练向量，扩充训练所使用的医学数据集获得了较好的命名实体识别效果；文献[12] 利用临床数据重新训练 BERT 预训练模型，并利用该预训练模型获得医学文本的嵌入式表示。然而重新训练领域预训练模型所需要收集的数据和耗费的资源是巨大的。对于更加细分的领域，更难以为其构建完全适合的预训练模型。此外，预训练模型在微调的过程中还普遍存在着灾难性遗忘[13] 的问题，即模型会遗忘从预训练语料中学习到的通用知识从而导致模型的泛化性降低。文献[14] 表明通过限制对 BERT 的模型参数更新程度，可以在一定程度上保留 BERT 的通用语言知识从而可以增强模型的泛化性。限制参数的更新也是以往工作中极少注意到的，例如文献[11 – 12] 都只是简单将预训练模型随微调过程进行更新，这样对于预训练模型会极易在参数更新的过程中遗忘其具有的通用语言知识造成对目标任务的过拟合。

为了解决当前预训练模型无法充分适应医疗领域的命名实体识别任务和在微调时易发生的灾难性遗忘问题，本文提出了一种基于回忆医学知识的预训练模型架构（RM-BERT）：在 BERT 预训练模型基础上融合医学的相关知识图谱从而构建了注入医学知识的 BERT 模型，并且使该融合医学知识图谱的 BERT 模型在微调过程中可以逐渐关注下游任务，通过这种方式可以使模型在更新参数以适应任务时避免损失太多通用的语言学知识，从而高效建模医学文本，实现医疗领域的命名实体识别。

1　回忆医学知识的预训练模型架构

本文提出的回忆医学知识的预训练模型架构见图 1，可分为两部分：一部分是融合医学知识的 BERT 模型，另一部分是为缓解融合医学知识的 BERT 模型在训练时的灾难性遗忘问题所使用的回忆学习策略。

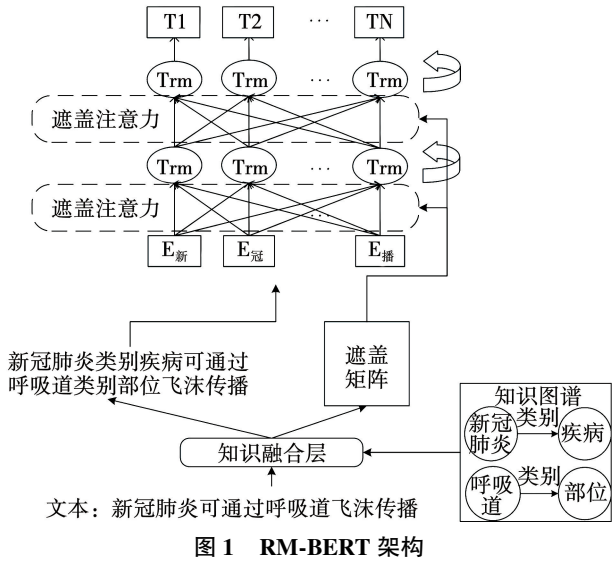

图 1　RM-BERT 架构

Fig. 1　Architecture of RM-BERT

架构的整体计算流程为：首先将输入文本与知识库中的知识三元组进行对比，在知识融合层匹配并引入相关知识，得到适合输入 BERT 模型的序列类型。然而直接融合知识三元组后的序列难以直接

表达原文本的含义,因此需要在知识融合层计算出文本的遮盖矩阵后注入 BERT 的自注意力计算过程。遮盖矩阵可以通过调整权重来控制 BERT 模型中的 Transformer[15] 捕捉原本语句和融入知识的特征。在利用预训练模型微调时,为了防止出现在微调过程中的灾难性遗忘问题,本文运用了一种回忆学习的策略使模型在微调时逐步关注目标任务。

本文提出了一种融合医学知识图谱的预训练模型架构:1)该架构融合医学知识图谱使预训练模型获得医疗领域文本的向量表示,以更好地应用在中文医疗命名实体识别任务中。2)使用回忆学习的策略来减弱融合知识的预训练模型在微调过程中过度更新参数而导致灾难性遗忘问题。

1.1 融合医学知识的 BERT 模型

1.1.1 BERT 预训练模型

本文使用 BERT 预训练模型[7] 作为融合医学知识的预训练模型。其通过在海量无标注语料上挖掘其中语义信息从而更好地建模文本表达。BERT 在许多任务上都有着较为出色的效果,其原因也来自于 Transformer 高效的建模自然语言文本能力。Transformer 本身是一种全连接的多头自注意力神经网络模型。多头自注意力机制的运算首先将输入向量序列 $\boldsymbol{H} \in \mathbb{R}^{L \times d_{\mathrm{m}}}$ 进行 3 种矩阵的线性变换,为自注意力机制提供所需的 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 矩阵,L 为序列的长度,d_{m} 为输入序列中每个单元的特征维度。因此该过程可由式(1)~(3)表示:

$$\boldsymbol{Q}_i = \boldsymbol{H}\boldsymbol{W}_i^Q \tag{1}$$

$$\boldsymbol{K}_i = \boldsymbol{H}\boldsymbol{W}_i^K \tag{2}$$

$$\boldsymbol{V}_i = \boldsymbol{H}\boldsymbol{W}_i^V \tag{3}$$

式中:$\boldsymbol{W}_i^Q \in \mathbb{R}^{d_{\mathrm{m}} \times d_{\mathrm{k}}}$,$\boldsymbol{W}_i^K \in \mathbb{R}^{d_{\mathrm{m}} \times d_{\mathrm{k}}}$,$\boldsymbol{W}_i^V \in \mathbb{R}^{d_{\mathrm{m}} \times d_{\mathrm{k}}}$ 为第 i 个语义空间进行线性变换的 3 个投影矩阵;d_{k} 为矩阵线性变换后的每个行向量的维度,是一个超参数。由此,每个头的注意力机制都可以由式(4)得到:

$$\boldsymbol{A}_i = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_i (K_i)^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V}_i \tag{4}$$

通过利用 softmax 函数计算求得每个语义空间的注意力矩阵。随后将每个语义空间求得的自注意力矩阵进行拼接后再进行一次矩阵变换可得到 Transformer 多头自注意力机制的计算结果。

1.1.2 融合知识图谱

目前较为公认的定义是:知识图谱可以看作是一种结构化的语义知识库[16],其采用“实体 - 关系 - 实体”三元组的形式来组成每一条知识。本文的工作是利用现有的知识图谱来增强 BERT 的语义理解能力,从而提升对医学电子病历的命名实体识别效果。具体的做法如下:

首先根据现有的知识图谱中的知识三元组对文本进行匹配,若匹配到相同的实体,则将该知识三元组与文本进行拼接插入。由于直接拼接插入知识会导致句意的不明确,因此在文本融合了知识图谱中的这些相关知识后需要补充额外的信息来方便模型理解文本的含义和知识。在知识融合层中,序列中的每个字符都标记了绝对位置和弹性位置。绝对位置指融合知识图谱后将句子按字符顺序依次标记的位置;弹性位置指未融合知识图谱时的字符顺序保留,拼接插入的知识三元组在相关实体的位置基础上按顺序依次标记。例如图 1 中的文本在融合知识后为“新冠肺炎类别疾病可通过呼吸道类别部位飞沫传播”,则其绝对位置和弹性位置的标记分别见图 2。

绝对位置:1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
融合文本:新 冠 肺 炎 类 别 疾 病 可 通 过 呼 吸 道 类 别 部 位 飞 沫 传 播
弹性位置:1 2 3 4 5 6 7 8 5 6 7 8 9 10 11 12 13 14 11 12 13 14

图 2 位置标记

Fig. 2 Position marker

弹性位置可以转化为 BERT 的位置嵌入,与文本的词嵌入和段落嵌入相加可以得到 BERT 的嵌入输入。这使得在 BERT 中的每一层 Transformer 计算注意力机制时,模型仍然可以按照正确的语句顺序来捕捉文本之间的关系。在融合后的文本中,增加的知识应当更关注于其涉及到的实体,对其他内容不应当有直接作用。例如图 2 中的融合文本中,“新冠肺炎”不应该与“类别部位”直接相关。因此需要将一些文本之间的注意力值进行遮盖。通过构建遮盖矩阵,使增加的知识仅与其涉及的相关实体直接作用。图 2 中例句的部分遮盖矩阵见图 3。

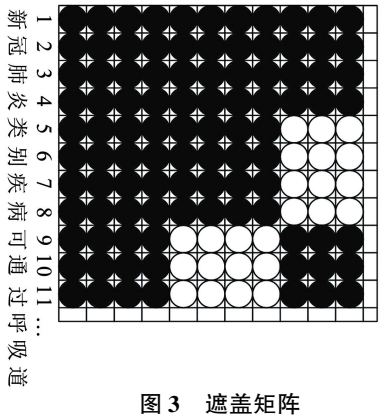

图 3 遮盖矩阵

Fig. 3 Masking matrix

遮盖矩阵是对称方阵，涂黑的格子代表该 2 个字符可以计算彼此之间的注意力，空白的格子表示该 2 个字符之间的注意力计算被遮盖了。例如“可通过”及之后的文本均遮盖对“类别疾病”的注意力，“类别疾病”也遮盖了对“可通过”及之后的文本的注意力。通过这种遮盖注意力的方式，每一层 Transformer 不仅可以对原文本的语义结构更加明确，并且可以将知识图谱中的语义信息融入相关的实体。由于 BERT 采用多层 Transformer 捕捉语义关系，因此在某一层 Transformer 将知识信息融合进入相关实体后可以在下一层更充分表达该实体的语义信息。层与层之间引入遮盖注意力的方法与原 Transformer 的注意力计算方法略有不同，见下式：

$$\boldsymbol{Q}_i^{l+1} = \boldsymbol{h}^l \boldsymbol{W}_i^{Q,l+1} \tag{5}$$

$$\boldsymbol{K}_i^{l+1} = \boldsymbol{h}^l \boldsymbol{W}_i^{K,l+1} \tag{6}$$

$$\boldsymbol{V}_i^{l+1} = \boldsymbol{h}^l \boldsymbol{W}_i^{V,l+1} \tag{7}$$

$$\boldsymbol{A}_i^{l+1} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_i^{l+1}(\boldsymbol{K}_i^{l+1})^{\mathrm{T}} + \boldsymbol{M}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V}_i^{l+1} \tag{8}$$

式中：$\boldsymbol{h}^l \in \mathbb{R}^{L\times d_{\mathrm{m}}}$ 为第 l 层 Transformer 的输出；$\boldsymbol{Q}_i^{l+1}$、$\boldsymbol{K}_i^{l+1}$和 $\boldsymbol{V}_i^{l+1}$ 分别为第 $l+1$ 层的第 i 个头的询问矩阵、键矩阵和值矩阵；$\boldsymbol{W}_i^{Q,l+1} \in \mathbb{R}^{d_{\mathrm{m}}\times d_{\mathrm{k}}}$、$\boldsymbol{W}_i^{K,l+1} \in \mathbb{R}^{d_{\mathrm{m}}\times d_{\mathrm{k}}}$和 $\boldsymbol{W}_i^{V,l+1} \in \mathbb{R}^{d_{\mathrm{m}}\times d_{\mathrm{k}}}$ 分别为相对应的转换矩阵。将其输入至第 $l+1$ 层的 Transformer 中后，计算第 $l+1$ 层中的每个头的注意力时，通过添加遮盖矩阵 $\boldsymbol{M} \in \mathbb{R}^{L\times L}$ 来达到控制遮盖注意力，L 为序列长度。$\boldsymbol{M}$ 中允许计算注意力的位置被设置为零，反之则为负无穷大。通过将遮盖矩阵融合进入 BERT 的每层 Transformer 中，可以使模型在融合知识图谱的知识基础上有效捕捉正常文本中的语义关系。

1.2　回忆医学知识策略

为了介绍回忆医学知识的策略，先引入多任务学习中的一些基本概念和设定。一般来讲，多任务学习可以分为目标任务和源任务。则其损失函数可看作

$$L_{\mathrm{M}} = \lambda L_{\mathrm{T}} + (1-\lambda) L_{\mathrm{S}} \tag{9}$$

$\lambda \in (0,1)$ 可以平衡两者的损失函数，由于源任务的损失函数 L_{S} 始终占据总损失函数的一部分，因此通过这样的方式可以有效避免灾难性遗忘这一问题。本文吸收了这种思想，将应用 BERT 进行迁移学习的过程分为两个阶段：一个是预训练阶段，这时训练的损失函数可以看作是 L_{S}；一个是微调阶段，这时训练的损失函数可以看作是 L_{T}。与多任务学习相似，可以通过设置混合损失函数来使预训练模型避免灾难性遗忘。不过由于任务最终的目的仍然是降低微调阶段的损失函数，也即 L_{T}。因此，本文受文献[17]启发，将调整损失函数的比例设置为随训练步数 t 变化的参数。公式为

$$L_{\mathrm{M}} = \lambda(t) L_{\mathrm{T}} + (1-\lambda(t)) L_{\mathrm{S}} \tag{10}$$

式中：$\lambda(t)$ 是随训练步数 t 变化而调整的比例系数，为了更好地平滑两种损失函数，$\lambda(t)$ 由 sigmoid 衰减函数计算得到。公式为

$$\lambda(t) = \frac{1}{1+\exp(-k(t-t_0))} \tag{11}$$

式中 k 和 t_0 为控制衰减速率和训练步数的超参数，通过设置这两个参数可以调整衰减函数的上升趋势，sigmoid 函数相较于线性函数和指数函数具有上升更加平滑并且易于通过两个超参数来控制曲线的优势。因此本文选择了该函数来达到控制模型回忆学习源任务的通用知识。

如图 4 所示，当 k 取正数时，该比例系数 $\lambda(t)$ 一开始较小，此时的损失函数更关注预训练的损失函数 L_{S}，随着训练步数的增加，损失函数愈加关注目标任务的损失函数 L_{T}。一般可以认为从 t_0 步开始，L_{T} 的权重开始逐步大于 L_{S} 的权重，通过这种方式可以让预训练模型在微调过程中不仅关注目标任务的损失函数，也能将预训练过程中的通用语义表示进行较好地保留。

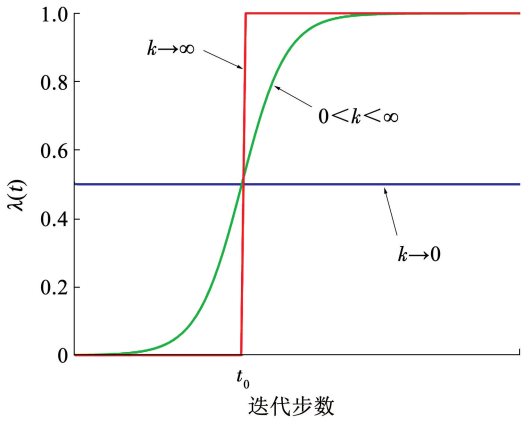

图 4　$\lambda(t)$ 函数图像

Fig. 4　Function $\lambda(t)$

由于重新获取预训练模型的训练数据是困难的，因此本文按照文献[18]对源任务的独立性假设和二次惩罚项来对 L_{S} 进行估计。对于源任务而言，其学习目标可以看作是优化给定源任务 D_{S} 时，模型参数 $\boldsymbol{\theta}$ 的负对数后验概率

$$L_{\mathrm{S}} = -\log p(\boldsymbol{\theta}|D_{\mathrm{S}}) \tag{12}$$

预训练的参数 $\boldsymbol{\theta}^*$ 某种程度上可以视为最小化损失函数的 $\boldsymbol{\theta}$ 值，此时的损失函数可以用式(13)估计：

$$-\log p(\boldsymbol{\theta}|D_S) \approx -\log p(\boldsymbol{\theta}^*|D_S) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\boldsymbol{H}(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*) \quad (13)$$

式中:$\boldsymbol{H}(\boldsymbol{\theta}^*)$是损失函数关于$\boldsymbol{\theta}$在$\boldsymbol{\theta}^*$处的海塞阵,由于$-\log p(\boldsymbol{\theta}^*|D_S)$为常数,因此在优化时可以将其忽略。根据文献[17-18]的推论,有式(14):

$$(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\boldsymbol{H}(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*) \approx NF\sum\nolimits_i(\boldsymbol{\theta}_i-\boldsymbol{\theta}_i^*)^2 \quad (14)$$

式中:N为源任务D_S中观测值的独立分布数量,F为给定源任务下的费雪信息。则优化L_S可以在不使用预训练数据的条件下由原参数得到,见式(15):

$$L_S = -\log p(\boldsymbol{\theta}|D_S) \approx \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\boldsymbol{H}(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*) \approx \frac{1}{2}NF\sum\nolimits_i(\boldsymbol{\theta}_i-\boldsymbol{\theta}_i^*)^2 = \frac{1}{2}\gamma\sum\nolimits_i(\boldsymbol{\theta}_i-\boldsymbol{\theta}_i^*)^2 \quad (15)$$

式中$\frac{1}{2}\gamma$可以看作是参数的二次惩罚项常数。由此可以得到整体的损失函数,通过优化该整体的损失函数即可达到控制模型回忆从源数据中学习到的通用语言知识。

2 实验及结果

实验所使用的计算机系统配置和主要程序版本如下:Windows10 操作系统,CPUi5-10400,内存16 GB,深度学习框架使用 Pytorch1.7,在单 NVIDIA GTX 3060 GPU 上训练模型。

2.1 数据来源与评价指标

实验使用公开的医疗病历数据集和通用领域数据集分别来自 2017 全国知识图谱与语义计算大会的医疗数据集[19]和微软研究院公布的公开数据集[20],本文将其分别命名为 Medical 数据集和 MSRA 数据集。标注的方式采用"BIO"的实体标注方式,即实体的第一个字符标记为"B-XX","XX"为其实体类型,实体的其余字符标注为"I-XX",其余字符标记为"O"。由于 Medical 数据集中已有相关工作[21]将数据集划分好,为了公平比较,本文直接沿用其对 Medical 数据集划分方法进行模型的测试。MSRA 数据集有公布的划分方式,本文直接沿袭这种划分方式。数据集所含实体情况见表 1 和表 2。

本文的评价指标按照命名实体识别通用的评价指标,即准确率(P),召回率(R),和 Micro-F1 值。

表 1 Medical 数据集信息

Tab. 1 Information of Medical dataset

Medical 数据集	训练集	开发集	测试集
身体部位	8 496	1 243	980
疾病症状	6 240	881	709
疾病实体	580	75	66
医学检查	7 566	1 113	867
治疗方式	853	97	98
全部	23 735	3 409	2 720

表 2 MSRA 数据集信息

Tab. 2 Information of MSRA dataset

MSRA 数据集	训练集	开发集	测试集
地点	16 235	1 846	3 527
机构	9 254	983	2 185
人物	8 100	883	1 862
全部	33 589	3 712	7 574

2.2 医疗领域与通用领域对比实验结果

本文以 BERT 模型为基本框架,采用融合知识的思路和回忆学习的策略来实现本文提出的方法。为了凸显本文提出方法的有效性,使用了多种现有的主流方法来对比本文提出方法在医学实体识别的效果。具体而言,本文对比了在复旦大学开源的 Word2Vec 静态预训练模型[22]作为词嵌入的基础上,使用双向长短期记忆(BiLSTM)模型和适应命名实体识别的改良 Transformer 模型(TENER)[23]作为编码器,条件随机场作为解码器的方法。由于本文在 BERT 模型的基础上进行了改进,因此还对比了原始的 BERT[7]模型以及同样在 BERT 的基础上进行改良的预训练模型,其中包括融合知识的 BERT 模型(K-BERT)[21]和增加了 N 元词组编码器的预训练模型(ZEN)[24]。本文对比实验中所对比的方法大多使用开源代码中的默认参数。在基于 BERT 的方法中,当数据集为 Medical 数据集时,实验设置学习率(learning rate)为3×10^{-5},批处理参数(batch size)为 10,最大句子长度(max sequence length)为 256,失活率(dropout)为 0.1,BERT 中每一层 Transformer 的 d_m 和 d_k 均设置为 768,总步数(step)为 14 000。对于 RM-BERT 方法,融合的知识图谱为医学概念知识图谱[21],设置回忆学习中的参数选取使用网格搜索法,最优时 k 为 0.01,t_0 为 1 000。实验结果展示了各个模型在 Medical 数据集的精确率、召回率和 Micro-F1。

表 3　Medical 数据集结果

Tab. 3　Results of Medical dataset　　%

模型	精确率	召回率	Micro-F1
BiLSTM	92.53	93.24	92.88
TENER	93.27	93.75	93.51
BERT	92.49	93.64	93.06
K-BERT	93.56	94.60	94.08
ZEN	93.56	94.49	94.02
RM-BERT	**94.07**	**94.49**	**94.28**

从表 3 中可以得到，在医学的命名实体识别任务上，只使用 BERT 预训练模型往往就可以得到较好的效果。然而由于 BERT 预训练模型使用的预训练语料与医学文本之间的隔阂较大，因此优势并不明显。使用一些高效的编码器如 TENER 所用的改进 Transformer 的效果甚至可以优于 BERT 的原始模型。通过将知识图谱融入预训练模型或者增加额外的特征编码器，可以在 BERT 预训练模型上得到更好的效果。K-BERT 引入知识图谱与 ZEN 增加额外的 N 元词组编码器取得了相近的效果。然而 K-BERT 并未考虑在迁移学习中的保留预训练模型中原本学习到的句法结构之类知识，ZEN 通过额外的 6 层 Transformer 结构使得模型参数量相比于 BERT 更为庞大。本文提出的方法在不引入额外的编码结构基础上，使模型回忆融合的医学知识图谱，提高了 BERT 本身对医学文本的表示能力和语义表达能力。融合知识图谱的过程中，本文采用弹性位置使模型在阅读融合知识图谱之后的文本时不仅保留原文本的顺序，并且将融入的知识图谱与对应实体也按顺序标记。使用该弹性位置可以将 BERT 预训练模型中每一层 Transformer 的自注意力机制计算进行有效指导，即利用遮盖矩阵使得融合的知识图谱信息更多作用在其对应的实体上，从而增强实体文本的特征表示。回忆机制中使用衰减函数使模型在训练时能够考虑源任务的损失函数，控制了预训练模型在目标任务上的参数更新，从而减缓了预训练模型在微调时的灾难性遗忘问题，增强了模型的泛化性。BERT 模型相较于浅层主流模型效果差这一问题也表明了使用通用领域文本做自监督训练导致模型难以在医疗领域充分发挥作用。通过使用本文提出的 RM-BERT 方法，将医疗知识图谱融入 BERT 模型后增强了 BERT 对医疗领域的建模能力，并使得其结果超越了浅层模型和主流深层模型。并且随着医学知识图谱的进一步扩展，该框架可以在更多医疗领域的信息抽取获得更具前景的效果提升。

为了证明本文提出框架的通用性和可扩展性，本文在进行 MSRA 数据集的实验时，更换 CnDbpedia[21] 作为知识图谱，学习率设置为 2×10^{-5}，批处理参数为 16，设置回忆学习中的参数 k 为 0.005，t_0 为 250。测试了在通用数据集 MSRA 上的结果，由于 MSRA 数据集已经有大量现有的研究成果，本文从一些文献中摘录了对应模型的识别效果，如 BiLSTM[25]、CAN-NER[26]、TENER[23]、ERNIE[27]。

表 4　MSRA 数据集结果

Tab. 4　Results of MSRA dataset　　%

模型	精确率	召回率	Micro-F1
BiLSTM *	92.97	90.80	91.87
CAN-NER *	93.53	92.42	92.97
TENER *	—	—	92.74
BERT	93.33	94.84	94.08
ERNIE *	—	—	93.80
K-BERT	93.73	95.18	94.45
ZEN	94.37	95.67	95.03
RM-BERT	94.15	95.36	94.75

由表 4 可得：由于 BERT 本身采用通用语料进行预训练，与 MSRA 数据集之间的隔阂并不大，基本上可以取得优于主流的浅层模型的效果。ERNIE 在预训练阶段注入知识对模型的改进在命名实体识别任务上的表现提升并不明显。K-BERT 在微调过程中融合知识图谱相较于 BERT 模型有一定的提升。本文提出的方法相较于 K-BERT 有进一步的提升。ZEN 模型采用额外的 6 层 Transformer 结构来编码 N 元词组获得了较高的提升，然而这一结构对于模型占用的空间也是不可忽视的。图 5 比较了基于预训练模型方法的模型大小，“bert-base”表示该模型为 BERT 原本的预训练模型，BERT 表示使用 BERT 预训练模型改造为进行命名实体识别的模型。由图 5 可得虽然本文提出方法在通用数据集上的效果稍微差于 ZEN 模型，然而 RM-BERT 模型相较于 ZEN 有小巧的优势。

综上，本文提出的方法在医学这类专业领域可以超过现有的浅层模型和深层模型的效果，在通用领域也优于浅层模型和大部分深层模型，在相同模型结构的类型中取得了最优的效果。不仅如此，本文提出的方法避免了重新预训练领域相关的 BERT 和增加额外的模型结构，这也使得在基于 BERT 预训练模型的基础上实现本文方法无需耗费过多的算力和显存。

2.3　消融实验结果

为了探究在基于回忆知识图谱的预训练模型架构中各部分所起的作用，本文进行了多组消融实验。

整体可分为:1)-MASK,去除遮盖矩阵,此时融入的知识图谱没有对应遮盖矩阵控制 BERT 中的注意力计算。2)-POS,去除文中使用的弹性位置编码,此时 BERT 只使用绝对位置编码的嵌入表达。3)-RECALL,去除回忆学习策略,此时模型的训练采用原始的 Adam 优化器优化。4)-KG,去除融入的知识图谱,则整体的架构也相应的没有遮盖矩阵和弹性位置编码引入自注意力机制中。消融实验的结果见图 6。

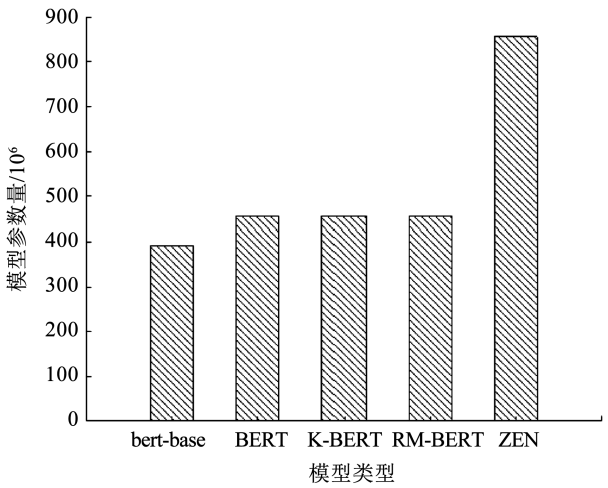

图 5 模型大小对比

Fig. 5 Comparison of model size

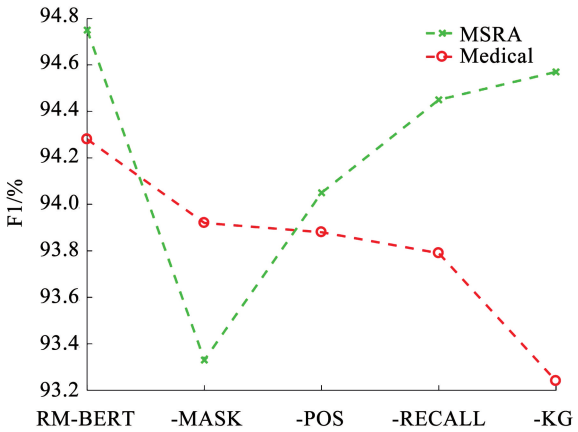

图 6 消融实验结果

Fig. 6 Ablation experiment results

由消融实验的结果可以得到通用领域和专业领域对于所提框架中的各部分依赖程度有所不同。去除遮盖矩阵后的 BERT 模型在捕捉句子内部中的注意力时不能得到有效的指导。由于本文将知识图谱融入原文本,文中所含有的实体应当与其融合的知识三元组直接相关,而对其他无关文本则不应该直接作用。因此去除遮盖矩阵使得 BERT 中的自注意力机制难以正确将知识图谱的信息与对应实体做注意力计算,从而损害了文本的建模,降低最终结果。对于 MSRA 数据集,去除遮盖矩阵会导致模型受影响程度更大,甚至低于去除知识图谱的效果,这说明此时引入的知识图谱在没有遮盖矩阵的指导下成为了影响模型的噪音。去除弹性位置编码而只使用绝对位置编码会导致融入知识图谱的句子不能很好地表达句子初始的语义从而影响模型的理解。在 Medical 数据集中,去除弹性位置编码的影响和去除遮盖矩阵影响相差不大,可以视为同等重要,MSRA 中则更依赖遮盖矩阵。回忆机制在 BERT 的微调过程使其在更新参数时考虑通用的语义表达,利用衰减函数动态调整目标任务和原模型的损失函数,缓解了迁移学习中所发生的灾难性遗忘问题。可以发现在医疗领域和通用领域,回忆知识这一策略都对模型有促进效果。对于医疗领域的命名实体识别任务,知识图谱所起到的作用是非常重要的,去除知识图谱后模型效果下降非常明显。然而对于通用领域的命名实体识别任务,知识图谱所发挥的作用并不像医疗领域中那样重要。这可能是由于 BERT 本身在通用语料上的训练已经获得了相当多的语义信息,知识图谱对于通用领域的语义提升并不十分明显。

2.4 实验分析

为了探究本文提出的回忆方法是否可以控制模型参数更新,本节比较了在 Medical 数据集中通过调整回忆算法中的 k 值来探究模型的参数更新状况。回忆知识的策略在模型更新的过程中注意比较更新后的模型参数与原 BERT 模型之间的参数。图 7所示为设置不同 k 值后,更新的 BERT 模型与原 BERT 模型的参数向量距离,也即对每一维度的参数作差取绝对值求和,t_0 设置为 1 000。

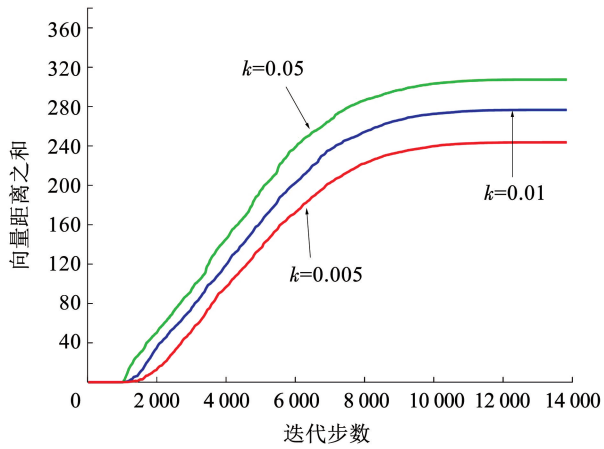

图 7 模型更新距离

Fig. 7 Distance of updating model

由图 7 可知,在刚开始的 t_0 步,损失函数中的衰减因子 $\lambda(t)<0.5$,模型的更新是微小的,此时模型更多学习通用语义表达;当更新的步数大于 t_0

时，损失函数中的衰减因子 $\lambda(t)>0.5$，模型参数的更新开始加快。通过设置不同的 k 值可以有效控制模型的参数更新程度，随着 k 值增大，衰减因子的函数曲线变得越陡峭，在步数大于 t_0 时目标任务的损失函数所占比例增加越快，模型在优化过程中利用 Adam 算法得到的更新越多。通过调整合适的 k 值参数可以得到模型最合适的更新程度。例如经过实验的探究，在 Medical 数据集上，当设置 $k=0.01$，$t_0=1\ 000$得到最优结果。当 k 值继续增大时，衰减因子对目标函数的损失函数限制减弱，此时利用 BERT 做迁移学习所引发的灾难性遗忘也越严重，模型也更容易在训练集上过拟合。因此使用回忆学习的策略需要将衰减函数的超参数 k 和 t_0 利用网格搜索法确定合适的值。并探究了 RM-BERT 在 Medical 数据集上各实体类型的识别效果，与原始 BERT 和 ZEN 的结果对比见表 5。

表 5　Medical 数据集实体类型结果对比

Tab. 5　Recognition results of entity types on Medical dataset　%

模型	身体部位	疾病症状	疾病实体	医学检查	治疗方式
BERT	88.48	97.11	71.53	96.86	90.91
ZEN	89.57	98.10	75.91	97.65	89.45
RM-BERT	89.70	98.09	78.83	98.05	90.00

由表 5 结果得到，3 个模型均在疾病症状上的识别最为有效，在疾病实体类型上的识别效果最差。RM-BERT 除了在疾病症状上效果比 ZEN 稍差，在其他类型的命名实体上都取得了更好的效果，尤其是在疾病实体效果上的提升最大，相较于 BERT 模型提升了 7.3%，相较于 ZEN 提升了 2.92%。由于疾病实体是标注实体中最少标注的类型，可以得出 RM-BERT 对标注量少的实体类型识别是有更显著帮助的。为了进一步分析本文提出方法的作用，对于实体数量最多的类别“身体部位”在训练集中进行了随机删除。通过调整不同的删除比例来观察减少训练集中实体对不同方法的影响。实验结果见图 8。

由于“身体部分”实体在训练集中数目较多，在删除 40% 时各个方法效果才有明显下降，本文在 40% 的删除概率基础上，以 10% 的概率依次增加删除概率并实验了相关方法。实验结果表明，本文提出的方法在训练实体减少时的识别效果下降更加平缓。在删除到 90% 的实体数目时也有着较其他两种方法更好的表现。通过表示学习的相关概念，本文提出的方法在相比于基础 BERT 模型在未增加额外向量维度的前提下取得了命名实体识别这一任务更好的效果，并且相较于增加向量维度的 ZEN 模型在医疗领域命名实体识别这一任务也取得了更好的效果。由于标签的解码输出采用相同的结构，可以认为本文提出的框架得到了更高效的医疗领域向量表示，从而获得了更好的命名实体识别效果。

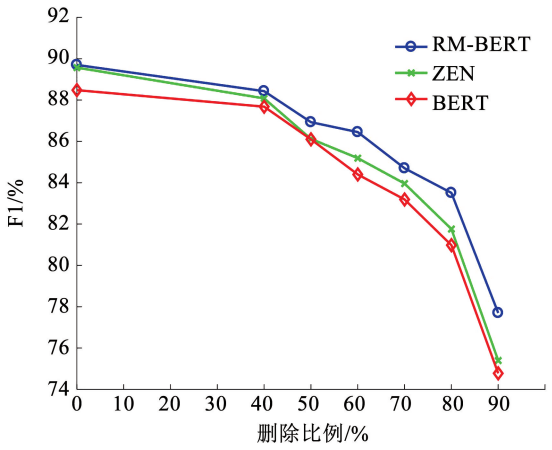

图 8　删减部分实体的识别效果对比

Fig. 8　Recognition effect of deleting parts of entities

3　结　论

本文提出了一种融合知识图谱的医疗领域命名实体识别架构，该架构利用弹性位置编码和遮盖矩阵有效地融合了知识图谱，并且使用回忆机制使模型关注通用的语义表达，使模型得到高效的医疗领域向量表示，最终得到医疗领域的实体标注类型。本文提出的方法在医疗数据集上较同类方法取得了最优的效果，在通用领域也有较好的表现。使用本文提出的模型可以在不增加额外编码结构和重新训练领域预训练模型的基础上有效识别专业领域内的命名实体。

参考文献

[1] WASSERMAN R C. Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research[J]. Academic Pediatrics, 2011, 11(4): 280. DOI: 10.1016/j.acap.2011.02.007

[2] DEMNER-FUSHMAN D, CHAPMAN W W, MCDONALD C J. What can natural language processing do for clinical decision support? [J]. Journal of Biomedical Informatics, 2009, 42(5): 760. DOI: 10.1016/j.jbi.2009.08.007

[3] LI Jing, SUN Aixin, HAN Jianglei, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(1): 50. DOI: 10.1109/TKDE.2020.2981314

[4] 刘知远，孙茂松，林衍凯，等. 知识表示学习研究进展[J]. 计算机研究与发展，2016，53(2)：247
LIU Zhiyuan, SUN Maosong, LIN Yankai, et al. Knowledge representation learning: a review[J]. Journal of Computer Research and Development, 2016, 53(2): 247. DOI: 10.7544/issn1000-1239.2016.20160020

[5] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2013: 3111. DOI: 10.5555/2999792.2999959

[6] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations [C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans: Association for Computational Linguistics, 2018: 2227. DOI: 10.18653/v1/N18-1202

[7] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019: 4171. DOI: 10.18653/v1/N19-1423

[8] 周大可, 宋荣, 杨欣. 结合双重注意力机制的遮挡感知行人检测[J]. 哈尔滨工业大学学报, 2021, 53(9): 156
ZHOU Dake, SONG Rong, YANG Xin. Occlusion-aware pedestrian detection combined with dual attention mechanism [J]. Journal of Harbin Institute of Technology, 2021, 53(9): 156. DOI: 10.11918/201904144

[9] 尹学振, 赵慧, 赵俊保, 等. 多神经网络协作的军事领域命名实体识别[J]. 清华大学学报(自然科学版), 2020, 60(8): 648
YIN Xuezhen, ZHAO Hui, ZHAO Junbao, et al. Multi-neural network collaboration for Chinese military named entity recognition [J]. Journal of Tsinghua University (Natural Science Edition), 2020, 60(8): 648. DOI: 10.16511/j.cnki.qhdxxb.2020.25.004

[10] GURURANGAN S, MARASOVI Ć A, SWAYAMDIPTA S, et al. Don't stop pretraining: adapt language models to domains and tasks [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2020: 8342. DOI: 10.18653/v1/2020.acl-main.740

[11] 罗凌, 杨志豪, 宋雅文, 等. 基于笔画ELMo和多任务学习的中文电子病历命名实体识别研究[J]. 计算机学报, 2020, 43(10): 1943
LUO Ling, YANG Zhihao, SONG Yawen, et al. Chinese clinical named entity recognition based on stroke ELMo and multi-task learning[J]. Chinese Journal of Computer, 2020, 43(10): 1943. DOI: 10.11897/SP.J.1016.2020.01943

[12] LI Xiangyang, ZHANG Huang, ZHOU Xiaohua. Chinese clinical named entity recognition with variant neural structures based on BERT methods[J]. Journal of Biomedical Informatics, 2020, 107: 103422. DOI: 10.1016/j.jbi.2020.103422

[13] MCCLOSKEY M, COHEN N J. Catastrophic interference in connectionist networks: the sequential learning problem [J]. Psychology of Learning and Motivation, 1989, 24: 109. DOI: 10.1016/S0079-7421(08)60536-8

[14] SUN Chi, QIU Xipeng, XU Yige, et al. How to fine-tune BERT for text classification? [C]//Proceedings of China National Conference on Chinese Computational Linguistics. Cham: Springer, 2019: 194

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017: 5998. DOI: 10.5555/3295222.3295349

[16] 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3): 582
LIU Qiao, LI Yang, DUAN Hong, et al. Knowledge graph construction techniques [J]. Journal of Computer Research and Development, 2016, 53(3): 582. DOI: 10.7544/issn1000-1239.2016.20148228

[17] CHEN Sanyuan, HOU Yutai, CUI Yiming, et al. Recall and learn: fine-tuning deep pretrained language models with less forgetting[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S. l.]: Association for Computational Linguistics, 2020: 7870. DOI: 10.18653/v1/2020.emnlp-main.634

[18] KIRKPATRICK J, PASCANU R, RABINOWITZ N, et al. Overcoming catastrophic forgetting in neural networks [J]. Proceedings of the National Academy of Sciences of the United States of America, 2016, 114(13): 3521. DOI: 10.1073/pnas.1611835114

[19] 中国中文信息学会语言与知识计算专业委员会. 电子病历命名实体识别[DS/OL]. [2022-03-24]. http://www.sigkg.cn/ccks2017/?page_id=51
Language and Knowledge Computing Professional Committee of Chinese Information Society of China. Electronic medical record named entity recognition[DS/OL]. [2022-03-24]. http://www.sigkg.cn/ccks2017/?page_id=51

[20] LEVOW G. The third international Chinese language processing bakeoff: word segmentation and named entity recognition [C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney: Association for Computational Linguistics, 2006: 108

[21] LIU Weijie, ZHOU Peng, ZHAO Zhe, et al. K-BERT: enabling language representation with knowledge graph [C]//Proceedings of The Thirty-Fourth AAAI Conference on Artificial Intelligence. Menlo Park: Association for the Advancement of Artificial Intelligence, 2019: 2901. DOI: 10.1609/aaai.v34i03.5681

[22] YAN Hang, QIU Xipeng, HUANG Xuanjing. A graph-based model for joint Chinese word segmentation and dependency parsing[C]// Transactions of the Association for Computational Linguistics. Cambridge: MIT Press, 2020: 78. DOI: 10.1162/tacl_a_00301

[23] YAN Hang, DENG Bocao, LI Xiaonan, et al. TENER: adapting transformer encoder for named entity recognition [Z]. arXiv pre-prints, 2019. DOI: 10.48550/arXiv.1911.04474

[24] DIAO Shizhe, BAI Jiaxin, SONG Yan, et al. ZEN: pre-training Chinese text encoder enhanced by N-gram representations [C]// Findings of the Association for Computational Linguistics: EMNLP 2020. [S. l.]: Association for Computational Linguistics, 2020: 4729. DOI: 10.18653/v1/2020.findings-emnlp.425

[25] ZHANG Yue, YANG Jie. Chinese NER using lattice LSTM[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018: 1554. DOI: 10.18653/v1/P18-1144

[26] ZHU Yuying, WANG Guoxin. CAN-NER: convolutional attention network for Chinese named entity recognition[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019: 3384. DOI: 10.18653/v1/N19-1342

[27] SUN Yu, WANG Shuohuan, LI Yukun, et al. ERNIE: enhanced representation through knowledge integration[Z]. arXiv pre-prints, 2019. DOI: 10.48550/arXiv.1904.09223

(编辑 苗秀芝)