计算机科学
COMPUTER SCIENCE

一种基于主动学习的文本实体与关系联合抽取方法

丁泓馨, 邹佩聂, 赵俊峰, 王亚沙

引用本文

丁泓馨, 邹佩聂, 赵俊峰, 王亚沙. 一种基于主动学习的文本实体与关系联合抽取方法[J]. 计算机科学, 2023, 50(10): 126-134.

DING Hongxin, ZOU Peinie, ZHAO Junfeng, WANG Yasha. Active Learning-based Text Entity and Relation Joint Extraction Method [J]. Computer Science, 2023, 50(10): 126-134.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300079

一种基于主动学习的文本实体与关系联合抽取方法

丁泓馨[1,2]　邹佩聂[1,3]　赵俊峰[1,2]　王亚沙[1,2]
1 北京大学计算机学院　北京 100871
2 高可信软件技术教育部重点实验室　北京 100871
3 北京大学软件与微电子学院　北京 102600
(dinghx@pku.edu.cn)

摘　要　非结构化文本数据中蕴含了大量有价值的知识,从中抽取出实体与关系形成结构化的知识,有助于知识图谱的构建,也可以为下游任务提供支持,具有广泛的应用前景。目前,实体与关系抽取问题多采用深度学习方法,但其模型的训练需要消耗大量标注数据,人工成本高,如何减少人工标注的工作量是当前研究的重点之一。主动学习是机器学习的领域之一,旨在通过选择最有价值的样本交予模型训练,在最大化模型性能增益的同时减少模型训练所需的数据量,其减少模型训练所需数据的潜力与深度学习数据贪婪的特性互补。因此,将主动学习应用到深度学习中的深度主动学习也是目前的研究热点。在上述背景下,使用深度主动学习进行实体与关系的联合抽取,将主动学习用于实体与关系抽取的深度学习模型的训练过程,在保持抽取模型性能的同时尽可能减少模型训练所需的人工标注数据。使用了一个基于统一标签空间、通过矩阵标注实现实体与关系联合抽取的深度学习模型,并在其基础上设计并实现了多种主动学习采样策略,在医疗领域的文本数据集和常用的实体与关系联合抽取数据集上验证了所提方法的有效性。对主动学习停止时机确定问题展开了研究,提出了根据模型训练损失曲线、模型在训练集上的性能、模型在预留数据上的预测稳定性来选择训练停止时机的方法,并通过实验研究了面向实际应用场景选取停止时机的方法。设计并实现了基于主动学习的文本实体与关系联合抽取的智能文本标注工具,可供用户对文本进行实体标注与关系标注,该工具实现了实体与关系抽取的深度学习模型与主动学习方法,可以最大程度地减少用户标注的工作量。

关键词:主动学习;知识抽取;命名实体识别;关系抽取;人机交互
中图法分类号　TP311

Active Learning-based Text Entity and Relation Joint Extraction Method

DING Hongxin[1,2],ZOU Peinie[1,3],ZHAO Junfeng[1,2] and WANG Yasha[1,2]
1 School of Computer Science,Peking University,Beijing 100871,China
2 Key Laboratory of High Confidence Software Technologies,Ministry of Education,Beijing 100871,China
3 School of Software & Microelectronics,Peking University,Beijing 102600,China

Abstract　Unstructured text data contains a large amount of valuable knowledge,entities and relations extracted from which can form structured knowledge and help to build knowledge graphs and support downstream tasks. There is a wide range of application prospects for entity and relation extraction. Currently,entity and relation extraction mostly use deep learning methods. However,the training of deep learning models consumes large amounts of annotated datasets,resulting in high labor cost. Therefore,how to reduce the workload of manual annotation is one of the focuses of research. Active learning is a subfield of machine learning,which aims to maximize a model's performance gain while annotating the fewest samples possible,by selecting the most valuable samples to be labeled and handed over to the model for training. Its potential to reduce training data complements the data-hungry nature of deep learning. Therefore,deep active learning that applies active learning in deep learning has become a hot research topic in entity and relation extraction. In the above context,using deep active learning for joint entity and relation extraction and appling active learning to the training process of the deep learning model to minimize the manual labeled data required for training while maintaining model performance,a deep learning model based on unified label space and matrix annotation for entity relation joint extraction is implemented and based on it,a variety of active learning query strategies are designed and implemented. The validity of the method is verified on text datasets and common entity and relation joint extraction datasets in the medical

到稿日期:2023-03-09　返修日期:2023-06-23
基金项目:国家自然科学基金(62172011);中央高校基本科研业务费专项资金
This work was supported by the National Natural Science Foundation of China(62172011) and Fundamental Research Funds for the Central Universities of Ministry of Education of China.
通信作者:赵俊峰(zhaojf@pku.edu.cn)

field. Several methods are proposed to select the stopping time of model training, including methods based on training loss curve of the model, model performance on the training set, and the prediction stability on reserved data. The method of selecting stop time for practical application scenario is studied by experiments. An intelligent text annotation tool based on active learning for joint extraction of entity and relation is designed and implemented, which allows users to annotate entities and relations in the text. The tool implements a deep learning model for entity and relation extraction and active learning methods to minimize the annotation workload of users.

Keywords Active learning, Knowledge extraction, Named entity recognition, Relation extraction, Human-machine interaction

1 引言

随着互联网、云计算、大数据等信息技术的发展,海量多源异构数据的处理与利用成为了待解决的热点问题。其中,知识图谱技术应运而生。知识图谱是节点和连边组成的图,是真实世界实体与关系的直观模型,这种知识表示方式适合理解、推理与解释,能以结构化的形式表示人类知识,有效解决跨系统、跨领域数据语义融合的问题。知识抽取是构建知识图谱的关键技术之一,其目的是从结构化、半结构化和非结构化数据中抽取知识实体及实体间的关系,用于知识图谱的构建或演化。利用结构化、半结构化数据进行知识图谱构建的技术较为成熟,而对非结构化数据的知识抽取与图谱构建,目前尚有诸多问题,是图谱构建技术的关键研究问题之一。

本文主要针对非结构化数据中的文本数据进行知识实体与关系的抽取。非结构化文本数据中蕴含了大量有价值的知识,从中抽取出实体与关系形成结构化的知识,有助于知识图谱的构建,也可以为下游任务提供支持,具有广泛的应用前景。例如,在医疗领域,随着医疗信息化的不断深入,电子病历数据、网络问诊信息、电子化的医学教材等数据越来越丰富,这些非结构化文本数据中蕴含了大量有价值的知识,人工识别和抽取知识成本高、效率低。因此,自动化的知识抽取技术,包括命名实体识别与关系抽取,得到了越来越多的关注。从文本中抽取出的实体与关系可以形成结构化的知识,为知识图谱的构建提供支持,帮助进行智能检索、知识推理等,有着广阔的应用前景。例如,从医学文本中抽取出的医疗实体及其关系可以用于构建医学知识图谱,在临床诊疗、疾病预后等多方面为医生和病人提供了广泛的帮助,发挥了重要作用。

知识抽取的相关技术经过多年的研究,从早期的利用规则工具抽取知识(包括基于硬模板的规则或基于软特性的规则),如 2005 年 Hanisch 等提出的基于规则识别基因蛋白质等实体的 ProMiner 系统[1],发展到从文本中根据句法、语义等信息构造一系列特征,并使用支持向量机(Support Vector Machine, SVM)、隐马尔可夫模型(Hidden Markov Model, HMM)、条件随机场(Conditional Random Field, CRF)等机器学习方法进行知识抽取,如 2012 年提出的基于 CRF 在化学领域进行命名实体识别的系统 ChemSpot[2],再到近期的利用深度学习模型抽取实体、关系。基于规则的方法需要根据领域知识与语言学知识人工制定规则或从大型语料库中挖掘规则,跨领域的迁移性较差且召回率低。而基于机器学习的方法需要复杂繁琐的特征工程,预处理工作复杂。目前,实体与关系抽取问题多采用深度学习方法,并取得了较高的准确率与召回率。

但是,深度学习模型的优良性能很大程度上得益于大型标注数据集。但在特定的专业领域,大型标注数据集是缺失的。在缺少训练数据的场景下,对文本进行实体与关系的自动抽取之前,需要人为标注一部分数据,用于训练深度学习模型。考虑到模型的质量高度依赖于数据的标注质量,专业领域文本的标注通常需要领域专家参与,例如医疗文本中蕴含大量医学专业知识,需要医学专家进行识别,人工标注成本高。如何在保证模型训练性能的同时减少需要标注的数据量,降低这一过程中的劳动成本,是一个亟需解决的重要问题。一种解决方案是使用主动学习训练深度学习模型,主动学习是机器学习的研究领域之一。主动学习期望通过选择当前最有价值的样本进行模型训练,来最大程度地改善模型,从而在达到预期训练效果的同时使用尽可能少的数据。

主动学习有助于减少深度学习模型训练所需的标注数据量,将主动学习应用到深度学习过程中的深度主动学习(Deep Active Learning)逐渐成为了研究热点。如何在非结构化文本实体与关系抽取问题上应用深度主动学习,设计并开发相应的实用工具,减小对文本进行实体与关系抽取时所需的高昂的人工标注成本,是本文研究的重点。本文的主要贡献如下:

1)构建了一个基于统一标签空间、通过矩阵标注实现实体与关系联合抽取的深度学习模型,且在模型中设计并实现了多种主动学习采样策略。

2)针对确定主动学习停止时机的问题展开了研究,提出了根据模型训练损失曲线、模型在训练集上的性能、模型在预留数据上的预测稳定性来选择训练停止时机的方法,并通过实验研究了面向实际应用场景选取停止时机的方法。

3)设计并实现了基于主动学习的文本实体与关系联合抽取的智能文本标注工具原型,并在 CBLUE 实验数据集上验证了本文方法的有效性。

本文第 2 节介绍了实体与关系联合抽取方法以及深度主动学习方法;第 3 节详细介绍了本文使用的实体与关系联合抽取模型以及提出的深度主动学习方法;第 4 节介绍了本文的实验设计及结果;第 5 节介绍了本文基于本文方法实现的智能文本标注工具;最后总结全文。

2 相关工作

2.1 文本实体与关系联合抽取

文本实体与关系联合抽取方法可分为流水线方法和联合抽取方法。流水线方法先进行实体抽取,随后将其两两配对,对抽取的实体进行关系分类,该方法的灵活性高,但存在误差积累、实体冗余与交互缺失的不足,模型性能仍有提升空间。联合抽取方法试图解决流水线方法的不足,利用实体和关系

之间存在的交互信息，来获得更好的性能。联合抽取方法可分为参数共享和统一标注两类。参数共享方法通过共享参数（如输入特征、隐状态…），联合实体抽取和关系抽取两个子模型，缓解了流水线方法中两个任务割裂与误差累积的问题。但在该类方法中，子模型独立解码，两者交互不强。统一标注方法则设计标注框架，将实体与关系抽取通过标注框架统一在一起进行建模，使用一个模型同时抽取文本中的实体与关系。该类方法能充分利用文本中实体与关系的交互信息，从而取得更优的性能。本文研究的重点是使用统一标注框架的联合抽取方法，下文将介绍这一方面的相关工作，重点介绍其设计的标注方式。

Zheng 等[3]首次提出使用统一标注框架进行实体与关系联合抽取，该工作将关系类别等信息融合到词语的 BIOS 标签中对句子进行标注，将实体与关系抽取问题转化为一个序列标注问题，并用端到端神经网络模型进行预测。但该方法不能直接识别实体类型，无法处理关系重叠问题，模型性能较差。

Wei 等[4]提出了一种级联标注方法，将关系视为从头实体映射到尾实体的函数，先预测得到头实体，对每一个头实体应用特定的关系函数判断有无对应的尾实体，从而实现实体与关系联合抽取。先使用头实体标注器两次标注句子，识别头实体的起始位置和结束位置，随后使用特定关系的尾实体标注器在给定头实体的情况下标注对应尾实体的起始与结束位置。但是，该方法无法识别实体类别，标签空间过大，标注过程繁琐。

Wang 等[5]设计了一种表格-序列编码器，使用序列标注执行命名实体识别任务，使用矩阵填充的标注方式处理关系抽取任务。使用 BIES-实体类别的标签进行序列标注识别实体，同时生成标签矩阵，其中第 i 行第 j 列位置的标签为对应词对的关系类别标签。该方法分别学习序列编码器和表格编码器，并使两个编码器互相交互，同时进行 NER 与 RE 任务的学习。该工作使用分离的编码器与标注任务处理实体识别与关系抽取，未能充分利用句子矩阵中实体与关系之间的交互与相互约束。

Wang 等[6]提出了一种基于统一标签空间的矩阵填充标注方法，句子矩阵中主对角线位置的正方形区域元素的标签即为对应位置词语的实体标签，非主对角线上位置的长方形区域元素的标签则表示对应实体之间的关系。在该标注方法下，属于同一实体的词语在矩阵中对应的行与列标签应该完全一致，解码时利用该特性识别实体范围，能取得比硬解码更鲁棒的效果。

2.2 深度主动学习方法

主动学习（Active Learning）是机器学习的一个研究领域。主动学习根据采样策略选择对当前模型训练最有价值的样本进行训练，以期在标注少量样本的同时使模型的性能增益最大化，由此减少模型性能达到预定标准时需要的数据量。其减少标注数据的目标与深度学习数据贪婪的特性互补，因此将主动学习方法应用到深度学习过程中的深度主动学习也是研究的热点之一。深度主动学习要求一次采样一批数据，下文将详细介绍这一方面的相关工作。

Shen 等[7]将深度主动学习应用在命名实体识别任务中，以减少训练所需的数据量。该工作使用轻量级的 CNN+CNN+LSTM 模型，来评估 4 种主动学习采样策略：最低置信度（Least Confidence，LC）方法、最大归一化对数概率（Maximum Normalized Log-Probability，MNLP）方法、基于分歧的贝叶斯主动学习（Bayesian Active Learning by Disagreement，BALD）和兼顾样本不确定性与代表性的方法。经实验证明，MNLP 方法在该问题上的表现最佳。

Zhdanov[8]希望在选择样本时综合考虑样本的信息量与样本的多样性，将模型对样本预测的不确定性作为样本信息量的衡量标注，先过滤出信息量最大的 $beta * k$ 个样本（$beta$ 为可调节的参数，k 为最终采样的样本量），随后在这些样本上进行 K 均值（K-Means）聚类，选取离聚类中心最近的 k 个样本作为最终的采样结果。

Ash 等[9]提出了一种基于多样梯度嵌入的批主动学习方法（Batch Active Learning by Diverse Gradient Embeddings，BADGE），该方法根据模型当前对样本的预测，计算样本针对模型最后一层参数的梯度嵌入向量，以此为基础通过 K-Means++初始化中心点的方法选择梯度最大且最多样的一批样本。该方法认为样本梯度的大小反映了样本对模型造成的影响，可以用来表示样本蕴含的信息量大小，而梯度嵌入的多样性也反映了样本的多样性。

大部分深度主动学习的研究工作都使用模型在验证集上的性能作为模型性能的评价指标，考察当模型达到预定性能标准时消耗的样本数量。这一指标的优势在于直观准确；劣势则是需要耗费额外的人力对验证集进行标注，不能充分利用标注数据。其他常见方法包括将模型对剩余样本的预测置信度作为指标，当其高于预先设置的阈值时停止训练。该方法存在的问题是，深度学习模型对训练样本的置信度往往很高，并不能反映模型实际的性能。另一种应用较广的方法则是选取一小批未标注数据，每次采样后在该批数据上进行预测，如果预测结果稳定，则认为模型已收敛，可以停止采样和训练。该方法存在的问题是小批数据难以代表数据分布，且常用的统计正负样本的衡量方式难以直接应用在实体与关系抽取任务上。

3 基于主动学习的文本实体与关系联合抽取方法

3.1 方法概述

本文针对非结构化文本实体与关系抽取任务，提出了一种人在回路的解决方案，该方案如图 1 所示。

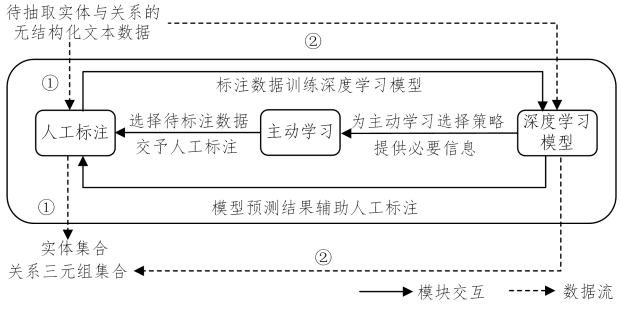


图 1 基于主动学习的实体与关系联合抽取方法的流程

Fig. 1 Process of active learning-based entity and relation extraction method

该方案使用主动学习方法将人工标注与深度学习相结合。对于预进行实体与关系抽取的无结构文本数据，人工标注其中部分数据并训练深度学习模型，使用模型对剩余数据进行自动抽取。而主动学习则根据深度学习模型提供的必要信息，通过选择对模型训练最有价值的样本来减少需要进行标注的数据量。

3.2 实体与关系联合抽取模型

问题的定义如下：模型的输入是一段文本 $S=x_1,x_2,\cdots,x_{|s|}$，实体标签集为Y_e，关系标签集为Y_r，模型将实体标签与关系标签置于统一的标签空间中得到标签集 $Y=Y_e\cup Y_r\cup\{\perp\}$。模型解码后的最终输出为实体列表与关系列表，每个实体需识别其在文本中的范围与实体类型，每个关系需识别其头实体、尾实体与关系类型。即：$entity:span=(start,end),type\in Y_e$，$relation:e1\in entites,e2\in entites,type\in Y_r$。

给定长度为$|s|$的句子，模型将构建一个规模为$|s|\times|s|$的矩阵 $\boldsymbol{T}$，并为矩阵中的每个位置$\boldsymbol{T}_{i,j}$预测标签。矩阵中主对角线位置的正方形区域元素的标签即为对应位置词语的实体标签，非主对角线上位置的长方形区域元素的标签则表示对应实体之间的关系。其他位置的标签则为$\perp$。

对于每个实体 e，矩阵中对应的位置$\boldsymbol{T}_{i,j}(i\in e.span,j\in e.span)$将被标注为其对应的实体类别 $type\in Y_e$。对于每个关系 r，矩阵中的位置$\boldsymbol{T}_{i,j}(i\in r.e1.span,j\in r.e2.span)$将被标注为其对应的关系类别 $type\in Y_r$。

例如，长度为 4 的句子$\{x_1,x_2,x_3,x_4\}$，$\{x_1,x_2\}$构成的词语属于 ENT1，x_4 属于 ENT2，且实体之间存在从 ENT1 到 ENT2 的关系，模型对一条文本的预测目标即为其对应的矩阵标注结果。标注样例如图 2 所示。文中使用的实体与关系联合抽取模型的结构如图 3 所示。

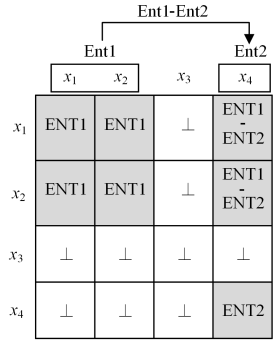


图 2 标注样例图

Fig. 2 Annotation example

输入的句子先通过预训练语言模型得到隐层表示：

$$\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_{|s|}\}=PLM(\{x_1,x_2,\cdots,x_{|s|}\})$$

得到句子的隐层表示后，模型使用双仿射注意力机制来捕获句子中任意词对之间的交互信息，从而对矩阵中的每个位置预测出标签。考虑到词语在参与关系时可能是头实体或尾实体，词语的隐层表示被分别输入一个 head mlp 和一个 tail mlp，得到头表示和尾表示，使得模型可以识别出当前词语是关系的头实体或尾实体，d 为该输出的表示向量的维度。

$$\boldsymbol{h}_i^{\text{head}}=MLP(\boldsymbol{h}_i)=MLP(\boldsymbol{h}_i),\boldsymbol{h}_i^{\text{head}}$$

$$\boldsymbol{h}_i^{\text{tail}}\in R^d$$

句子矩阵中$\boldsymbol{T}_{ij}$代表词 i 与词 j 形成的词对，预测该位置标签时，将词 i 的头表示和词 j 的尾表示输入双仿射模型计算得到概率得分，得分通过 Softmax 层得到模型预测的概率分布。

for each $i,j\in(0,|s|)$

$$\boldsymbol{h}_1=\boldsymbol{h}_i^{\text{head}},\boldsymbol{h}_2=\boldsymbol{h}_j^{\text{tail}}$$

$$\boldsymbol{g}_{i,j}=\boldsymbol{h}_1^{\mathrm{T}}\boldsymbol{U}_1\boldsymbol{h}_2+\boldsymbol{U}_2(\boldsymbol{h}_1\oplus\boldsymbol{h}_2)+\boldsymbol{b}$$

$$\boldsymbol{P}_{i,j}=\text{softmax}(\boldsymbol{g}_{ij}),\boldsymbol{P}\in R^{|S|\times|S|\times|y|}$$

由此，模型预测了矩阵中每个位置在标签空间中的概率分布。

模型的预测损失值包括矩阵中每个位置与正确标签之间的交叉熵损失。

$$\mathscr{L}_{\text{entry}}=-\frac{1}{|s|^2}\sum_{i=1}^{|s|}\sum_{j=1}^{|s|}\log P(\boldsymbol{y}_{i,j}=y_{i,j}\,|\,s)$$

此外，预测损失值还包括正则约束项对称损失与暗指损失，以更充分地利用实体与关系的信息。对称损失约束实体区域标签与对称关系的区域标签关于对角线对称。

$$\mathscr{L}_{\text{symmetric}}=\frac{1}{|s|^2}\sum_{i=1}^{|s|}\sum_{j=1}^{|s|}\sum_{t\in Y_{sym}}|\boldsymbol{P}_{i,jt}-\boldsymbol{P}_{j,i,t}|$$

暗指损失则约束关系概率不能低于对应位置存在实体的概率，若要将$\boldsymbol{T}_{ij}$识别为关系 t，则x_i是实体的概率和x_j是实体的概率需大于概率$\boldsymbol{P}_{i,j,t}$。

$$\mathscr{L}_{\text{implication}}=\frac{1}{|s|}\sum_{i=1}^{|s|}\left[\max_{t\in Y_r}\{\boldsymbol{P}_{i,:,t},\boldsymbol{P}_{:,i,t}\}-\max_{t\in Y_e}\{\boldsymbol{P}_{i,i,t}\}\right]$$

模型训练时的优化目标是$\mathscr{L}_{\text{entry}}+\mathscr{L}_{\text{symmetric}}+\mathscr{L}_{\text{implication}}$。解码时，采用一种巧妙的 Span decoding 的解码方式。从图 3 中可以观察到，矩阵中属于同一个实体的词语所对应的行或列是完全相同的，基于该观察，可以通过计算行间、列间的距离来进行区域的分割划分，获得句子中实体的文本范围(text span)，再进一步计算矩阵中两实体范围构成的正方形、长方形区域中标签的平均得分，取得分最高的标签类型作为对应实体、关系的类型。

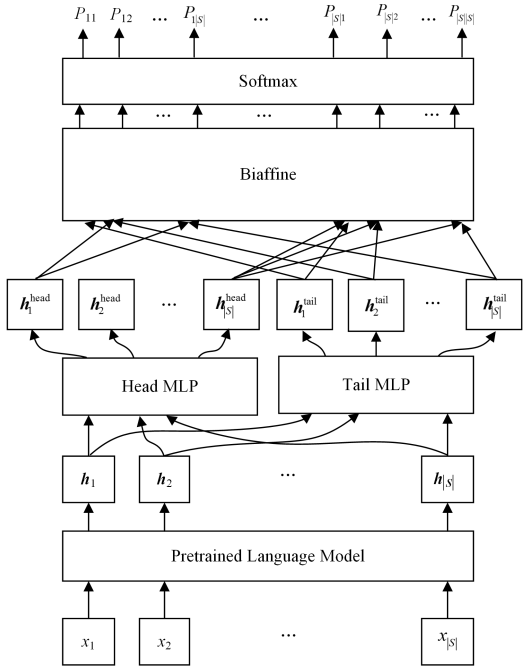


图 3 模型结构图

Fig. 3 Model structure

3.3 主动学习采样策略

主动学习采样策略在每次训练模型前选择需要加入训练集的样本，好的采样策略将选择对模型训练最有价值的一批样本，使得模型每次训练都获得最大的性能提升，进而减少采样与训练的次数，这是主动学习减少训练所需数据量的关键。

主动学习算法的伪代码如算法 1 所示。

算法 1 主动学习算法

输入：神经网络 f(x;θ)，未标注样本 U，采样策略 Q，采样大小 K，每轮训练 epoch 数 I，停止条件 C

输出：训练完成的网络 f(x;θ)

1. Labeled training set S←K examples selected at random from U and get their queried labels.
2. Train an initial model f(x;θ) on S for I epochs.
3. Do:
4. S_t←K examples from U using query strategy Q and get their queried labels.
5. S←S∪S_t
6. Train the model f(x;θ) on S for I epochs
7. Until condition C is satisfied or unlabeled samples U is exhausted

主动学习算法中的采样策略 Q 可根据具体应用场景确定。为了选择最适合实体与关系抽取任务与上述深度学习模型的采样策略，最大程度地减少人工标注量，本文设计并实验了多种主动学习的样本选择策略，包括基于不确定性的策略、基于样本多样性的策略以及混合策略。其中基于不确定性的策略包括最小置信度(Least Confidence)方法、基于边界(Margin)的方法、基于熵(Entropy)的方法；基于样本多样性的策略是 K-均值(K-means)方法；混合策略包括基于梯度的 badge 方法和基于熵的 K 均值(Entropy K-means)方法。

1)基于模型不确定性的策略

这一类策略认为，模型预测不确定的样本蕴含较多模型未学习到的特征，可以最大程度地改善模型性能。这些方法使用不同的指标度量模型对样本预测的不确定性，计算每个样本相应的不确定性得分，对得分进行排序后选取其中不确定性得分最高的一批样本。在该问题中，一个样本是一个句子，我们定义该句子为 s，句子长度为 $|s|$，标签空间为 Y，则深度学习模型为句子矩阵中的每个位置预测其在标签空间中的概率分布向量，模型的输出是矩阵 $\boldsymbol{P}\in R^{|s|\times|s|\times|Y|}$，$\boldsymbol{P}_{i,j,t}$ 表示将矩阵中位置 ij 的标签预测为 t 的概率。我们根据该矩阵计算如下几种不确定性度量得分。

(1)Least confidence：以模型对样本标签预测的最大概率反映模型对样本预测的不确定性，最大概率越低，其不确定性得分就越高，计算式如下：

$$LC\_Score_s=1-\frac{1}{|s|^2}\sum_{i=1}^{|s|}\sum_{j=1}^{|s|}\max_{t\in Y}\{\boldsymbol{P}_{i,j,t}\}$$

(2)Margin：以模型对样本预测的概率分布中最大概率和次大概率的差值表示模型预测的不确定性，该差值越小，代表模型越不容易分辨当前样本的标签，不确定性得分越高，计算式如下：

$$Margin\_Score_s=1-\frac{1}{|s|^2}\sum_{i=1}^{|s|}\sum_{j=1}^{|s|}\max_{t_1\in Y}\{\boldsymbol{P}_{i,j,t_1}\}-\max_{t_2\in Y-\{t_1\}}\boldsymbol{P}_{i,j,t_2}$$

(3)Entropy：以模型对样本预测的概率分布的信息熵表示模型预测的不确定性，信息熵越高，其不确定性得分就越高，计算式如下：

$$Entropy\_Score_s=-\frac{1}{|s|^2}\sum_{i=1}^{|s|}\sum_{j=1}^{|s|}\sum_{t\in Y}\boldsymbol{P}_{i,j,t}\log\boldsymbol{P}_{i,j,t}$$

在得到模型对所有未标注样本的不确定性得分后，从中选取得分最高的一批进行标注与训练。

2)基于样本代表性与多样性的策略

K-means：该方法是基于样本代表性与多样性的策略，该策略希望选择一批能尽可能代表其他无标注样本且彼此之间相似度较小的样本，使得模型可以学习到数据的典型特征，且获得的冗余信息较少。对于输入的一个句子，模型使用预训练语言模型获得每个词的隐藏表示，随后通过两个多层感知机获得每个词的头表示$\boldsymbol{h}_i^{\text{head}}$和尾表示$\boldsymbol{h}_i^{\text{tail}}$。我们拼接每个词的头尾表示，并对句子长度取平均，将其作为一个样本的表示向量。

$$\boldsymbol{h}_s=\frac{1}{|s|}\sum_{i=1}^{|s|}(\boldsymbol{h}_i^{\text{head}}\oplus\boldsymbol{h}_i^{\text{tail}})$$

获得所有样本的表示向量后，以欧氏距离为度量，使用 K-means 算法对其进行采样数聚类。我们认为聚类中心可以代表一个聚类的样本，并且聚类中心往往彼此间距离较远，具有一定多样性，因此其能更好地覆盖样本分布。故该方法选取距离聚类中心最近的一批样本加入训练集。

3)混合策略

(1)BADGE：沿用文献[9]中提出的主动学习方法。如相关文献所述，该方法根据模型当前的预测结果，计算样本针对模型最后一层产生的梯度估计，梯度越大，说明该样本对模型的影响越大，其蕴含的信息量可能越多。同时，为了兼顾所选样本的多样性，在梯度表示上使用 K-means++初始化算法，来获得彼此间距离最远的一批样本。

(2)EntropyKmeans：参考文献[8]，希望兼顾模型对样本预测的不确定性与样本的多样性，避免选择一批不确定但相似的样本，造成信息冗余。因此，使用上述 Entropy Score 的计算方法，先选择信息熵最大的 k(采样数) * $prefilter\ number$ 个样本，在其上进行 K-means 聚类，选择距离聚类中心最近的一批样本。

3.4 主动学习停止时机选择方法

主动学习期望减少人工标注的工作量，因此机器应当自动识别出何时模型已经获得相对较好的性能，可以停止进一步的标注与训练，使用深度学习模型对剩余的未标注文本进行自动抽取。

在实验中，我们可以根据模型在验证集上的预测准确度来衡量模型性能。但在真实场景中，有真实标签的验证集往往是难以获取的，或需额外进行人工标注。因此，需设计其他指标来衡量模型性能是否已达到预期，可以停止训练。本文设计了以下 3 种衡量指标。

1)基于模型在训练集上的准确度

在主动学习的场景中，虽然不存在验证集，但已标注的训练集可以充当验证集的角色，当模型在训练集上的训练准确度达到预先设定的阈值时停止训练。

2)基于模型损失值曲线

深度学习模型的训练效果、模型是否稳定收敛,一般可以通过模型的 loss 下降曲线和预测结果得出结论。当模型的损失值曲线较为平缓,下降的速率小于某个阈值时,我们认为模型已经得到较为充分的训练,可以停止进一步的采样与训练。我们采用两种方法来判定模型的损失值曲线是否已平稳。

(1)近似计算损失值曲线拐点。在训练过程中,模型损失值先迅速下降,后下降趋势变得平缓,将损失值曲线开始变缓的位置称为曲线的拐点,并认为拐点后损失值曲线已平稳。精确计算拐点需要对曲线进行二次求导,我们使用一种简洁的近似计算方法,连接曲线头尾形成直线,计算曲线上各点到该直线的垂直距离,垂直距离最大的点可以近似认为是曲线的拐点。可以选择 loss 曲线的拐点作为主动学习训练流程停止的时间点。

(2)近似计算损失值下降速率。采用差分近似计算模型损失值的下降速率,使用当前损失值减去上一次迭代的损失值,近似为损失值的下降速率。当该下降速率小于某一事先给定的阈值时,停止主动学习的采样与训练。考虑到模型的损失值与具体的预测任务及数据集有关,设置绝对阈值是不可取的,我们可以设置相对阈值,如两次迭代中模型的损失值下降小于模型初始损失值的 1%时停止训练。

3)基于模型预测稳定性

选取少量数据作为 stop batch,当模型在 stop batch 上的预测基本稳定时停止训练。考虑到一个样本是一个句子,模型预测结果是其中的实体与关系,在每次迭代后获得模型从 stop batch 中抽取出的实体与关系,并统计当前模型的抽取结果与上一次迭代的抽取结果的交集与并集,计算交集大小与并集大小的比值。如果该比值连续多次迭代均大于预设的阈值,则认为模型的预测结果已经稳定,可以停止进一步的采样与训练。

4 实验设计与结果分析抽取方法

4.1 实验设计

基于本文方法,本文主要使用主动学习结合深度学习对文本进行实体与关系抽取。由前述分析可知,为了达到最大限度减少人工标注量的目的,需要选择合适的主动学习采样策略和有效的主动学习停止时机选取方法。本实验的目的是验证第 3 节中提出的相关方法的有效性,并从中选取最优方法。

对于主动学习采样策略,我们需要验证设计的策略能否有效减少模型训练所需的数据量,即使用采样策略进行有选择的采样,模型只需使用全量数据的小部分进行训练即可获得较好的性能。我们计划记录主动学习采样过程中模型性能的变化,对于不同的主动学习采样策略,在采样了相同次数时,模型性能越好,说明该采样策略越有效,能在使用了相同数据量的情况下更好地改善模型性能。此外,我们也设定模型预期性能,比较使用不同主动学习方法训练模型时达到同一性能所需的采样次数、消耗的数据量。

对于主动学习停止时机的确定,我们需要验证设计的衡量指标能否很好地反映模型性能并指示主动学习过程在恰当的时机停止。我们计划统计不同衡量指标在主动学习采样训练过程中的变化,查看当所使用的指标达到阈值时模型的性能是否符合预期。

4.2 实验验证

4.2.1 实验数据

考虑到通用领域已有成熟的大规模数据集,本文主要关注在领域文本上进行实体与关系抽取的场景。在这种场景下,标注数据集较难获得,文本标注的成本高,智能标注工具能发挥更大的作用。本文选择在医疗领域文本上进行实验,使用的数据集为中文医疗信息处理挑战榜 CBLUE(Chinese Biomedical Language Understanding Evaluation)[10] 提供的两个数据集,分别是中文医学命名实体识别的 CMeEE 数据集和中文医学文本实体与关系抽取的 CMeIE 数据集。

CMeEE 数据集[11] 包括训练集数据 15 000 条,验证集数据 5 000 条。该数据集中待识别的医学文本命名实体共有 9 类,包括疾病、临床表现、药物、医疗设备、医疗程序、身体部位、医学检验项目、微生物类和科室。任务要求自动识别句子中的实体,输出实体集合。

CMeIE 数据集[12] 包括训练集数据 14 339 条,验证集数据 3 585 条,数据集中共含有近 7.5 万个三元组。CMeIE 给出了 53 种定义好的 schema,定义了关系以及其对应的头实体和尾实体的类别,例如("subject_type":"疾病","predicate":"药物治疗","object_type":"药物")。任务要求自动对句子进行分析,输出句子中所有满足 schema 约束的 SPO 三元组知识 $Triples=[(S1,P1,O1),(S2,P2,O2),\cdots]$。

对于命名实体识别任务 CMeEE,采用严格 Micro-F1 作为评价指标,要求同时精确预测出实体的起始、结束下标以及实体类型。

对于实体与关系抽取任务 CMeIE,采用严格 Micro-F1 作为评测指标,要求同时精确预测出关系三元组的头实体、头实体类型、尾实体、尾实体类型和关系类型。

4.2.2 实验过程

本文使用 PyTorch 进行神经网络的搭建和训练,使用的预训练语言模型为 chinese-bert-wwm-ext。本文沿用 Wang 等的工作,将模型 MLP 层的隐层大小设置为 150,并使用 GELU 作为激活函数。训练模型时,使用 AdamW 作为模型优化器,其中 $beta1=0.9$,$beta2=0.9$。学习率为 5×10^{-5},权重衰减为 1×10^{-5},使用带有 warmup 的线性学习率调整器,由于主动学习的训练方式不同于原模型的训练方式,我们在每次采样后重新设定学习率 scheduler。在训练时,batch 大小为 32,主动学习一次采样 6 个 batch 的数据,即 192 条文本。采用增量式训练,每次采样后,在所有已标注数据上继续训练模型。由于模型不能处理实体嵌套的情形,因此模型在数据加载阶段会将数据集中具有嵌套实体的句子剔除,此外模型对输入句子的长度有限制,过长的句子也将被剔除。剔除后,主动学习实验在 CMeEE 数据集上共进行 69 次采样,在 CMeEE 数据集上进行 71 次采样。

实验过程中,每次采样和训练后我们都使用验证集对模型进行验证,记录其 F1 score,以直观地衡量模型在主动学习

训练过程中的性能变化。此外，为了验证主动学习停止时机确定问题中设计的指标，每次采样和训练后也记录模型在一次采样训练过程中的平均损失值、模型在所有训练数据上的F1 score和模型在一个batch的预留数据上的训练结果。

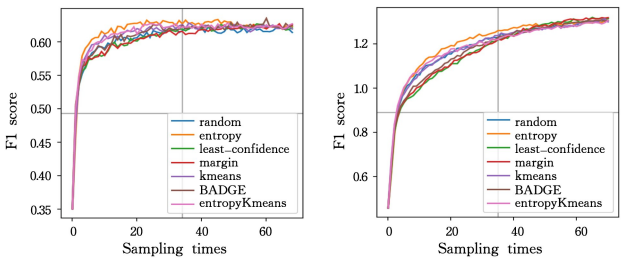
4.3 实验结果

4.3.1 主动学习策略选择

为了体现主动学习减少训练所需数据量的作用，首先使用全量数据对模型进行训练。采用early stop机制，模型训练200个epoch之后，若模型性能连续30个epoch没有超过最佳性能便停止训练。在CMeEE数据集上，模型在验证集上取得的最优F1 score为64.77%，在CMeIE数据集上，模型在验证集上的最优结果是实体预测的F1 score，为75.33%，关系预测的F1 score为59.32%。

根据4.1节中的实验设计，对3.2节中叙述的主动学习算法进行全面的实验与比较分析。使用随机采样策略作为实验的基线方法，该方法每次在未标注数据集上随机采样 k 个样本交予模型用于训练，有效的采样策略应该取得比随机采样更优的结果。实验结果如图4所示。

从不同采样策略下模型在验证集上的F1 score曲线图可以看出，在主动学习过程中，采样了70%的数据之前，在两个数据集上均是基于Entropy的采样策略最有效。在采样了相同的样本量时，使用Entropy采样策略训练的模型性能明显优于包括随机采样在内的其余方法，其模型也是最快达到较为优良的性能的。在采样了70%的样本后，各种方法的训练结果的性能相差较小。在CMeEE数据集上 K-means与EntropyKmeans的前期性能略高于Random方法，在CMeIE数据集上与Random方法的性能相近。Least confidence和Margin方法在采样前期性能较差，在CMeIE数据集上采样结束时性能较优，而BADGE方法则不具备明显的优越性。



(a)CMeEE epochs:50 (b)CMeIE epochs:50

图4 不同采样策略下的模型F1score曲线

Fig. 4 F1 score curve under different sampling strategies

此外，我们选取使用全量数据训练的模型性能的90%和95%为目标，考察使用主动学习方法达到训练目标时需要进行的采样次数。需要的采样次数越少，说明该主动学习采样策略越有效，能更好地减少模型训练所需的数据量。

如表1和表2所列，基于Entropy的采样策略在两个任务上达到两个预设目标时所用的采样次数，即需要消耗的数据量都是最少的。在CMeEE任务上，该方法使用7.2%的数据即可达到全量数据训练效果的90%，使用14.5%的数据量则可达到全量数据训练效果的95%；在更为复杂的实体与关系联合抽取任务上，使用33.8%的数据量可以达到全量数据训练效果的90%，使用61.9%的数据可达到全量数据训练效果的95%。

表1 各采样策略达到全量数据训练性能90%所需的采样次数

Table 1 Number of queries required for each sampling strategy to achieve 90% of the training performance of using full amount of data

Dataset	Random	Entropy	Least confidence	Margin	K-means	BADGE	Entropy Kmeans
CMeEE	7	5	12	10	6	8	6
CMeIE	30	24	34	36	30	34	34

表2 各采样策略达到全量数据训练性能95%所需的采样次数

Table 2 Number of queries required for each sampling strategy to achieve 95% of the training performance of using full amount of data

Dataset	Random	Entropy	Least confidence	Margin	K-means	BADGE	Entropy Kmeans
CMeEE	26	10	29	31	18	23	18
CMeIE	53	44	48	48	55	53	55

综上所述，我们发现基于Entropy的采样策略在本文模型与任务上有着最优越的性能，能在使用最少训练数据的情况下取得最优的模型性能，符合减少人工标注量的要求，适于应用在文本标注的工具中。

4.3.2 主动学习停止时机

本文使用在采样中表现最佳的基于entropy的采样策略进行确定主动学习停止时机的相关实验。实验验证了第3节中提出的4种停止策略：基于loss曲线拐点的策略、基于loss曲线下降速率的策略、基于stop batch的策略、基于模型在训练集上的准确率的策略。

1)基于Loss曲线拐点的策略。该策略需要获得完整的loss曲线，即模型需要进行全部采样，这与主动学习的目标是相悖的。此外，计算得到的拐点对于同一条loss曲线是固定的，不具备根据用户需求进行调整的能力。该方法不可行。

2)基于Loss曲线下降速率的策略。该策略设定当两次迭代损失值的差值小于其起始损失值的1%时可以停止训练。在CMeEE任务中，主动学习在第6次采样后停止；在CMeIE任务中，主动学习在第15次采样后停止训练。可以发现，基于loss曲线下降速率的方法的阈值设置较为困难，难以提前选定合适的阈值，在实际应用中不可行。

3)基于Stop batch的策略。经实验发现，stop batch的方法在本文模型上是不可行的。当一次采样训练的epoch为50时，模型在stop batch上的预测结果波动较大，在采样接近结束时才能保持稳定，我们认为这一结果是因实体与关系联合抽取模型较为复杂、训练任务难度大，主动学习一次采样的样本量较小且训练次数多造成的，此种训练方式容易导致模型在当前已采样数据上过拟合，造成模型波动较大。而当一次采样后训练的epoch为10时，模型在stop batch上的预测结果很快就稳定且始终保持不变。这两种情况下，该方法都无法探测出模型达到良好的效果，可以停止训练的时机。此外，stop batch的数据量太小，也不能代表整个数据的真实分布，以小批量数据的训练结果为标准难以衡量模型的真实情况。

4)基于模型在训练集上的准确率的策略。设定当模型在

训练集上的 F1 score 达到 99%时停止训练，在 CMeEE 任务中，主动学习在第 6 次采样后停止，此时模型在验证集上的性能已超过全量数据训练性能的 90%；在 CMeIE 任务中，主动学习在第 21 次采样后停止，与模型达到全量数据训练性能 90%所需的 24 次采样数基本相符。在实际应用中，可以根据该实验经验进行阈值的选取。

综上所述，使用模型在训练集上的准确率作为衡量指标，能较好地在模型性能基本符合预期时停止。此外，该指标直观且意义明确，在实际应用中，用户可以根据需求提高或降低阈值，以满足其对标注工作量和模型准确度的取舍，适于应用在文本标注工具中。

5 智能文本标注工具

基于上文介绍的 UniRe 模型与多种主动学习方法，本文设计并实现了一个搭载了主动学习采样引擎与深度学习模型的文本标注工具。

我们使用 Vue 作为工具前端的框架，使用 Django 作为工具的后端框架，以 sqlite 作为数据库。模型部分则使用 pytorch 进行深度学习模型构建与主动学习策略的实现，并使用 zmq 实现模型与工具后端之间的通信。

文本标注工具主要包括以下几个主要界面，如图 5、图 6 所示。

图 5 实体标注界面

Fig. 5 Entity annotation interface

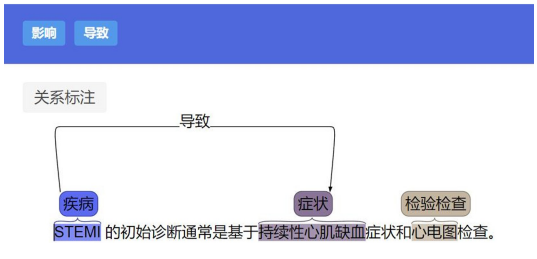


图 6 关系标注界面

Fig. 6 Relation annotation interface

1)数据与设置界面

用户可以上传待标注文本数据、本体模型与实体词典。用户可以对标注的实体标签、关系标签、训练的参数与主动学习方法等进行编辑与设置，也可以查看当前对数据集标注的统计状态，并在标注结束后导出数据。

2)智能标注界面

该部分是工具的核心功能，在这一部分中用户对文本进行实体与关系标注。该界面左侧列出了待标注的文本，界面中央为对文本进行标注的部分，标注部分分为用户标注与 AI 智能推荐两部分。

图 5 给出了实体标注界面，用户手动标注实体时，使用鼠标划出实体的范围，点选句子上方的候选标签进行标注。

在 AI 智能推荐部分，工具使用下划线标明 AI 推荐的实体标注。用户通过智能推荐进行标注时，只需单击下划线标明的区域，就可弹出提示框，显示推荐的标签，通过点选即可在上方用户标注部分完成标注。

在标注了一条文本的实体标注后，可以点击标注关系的按钮进入关系标注界面，关系标注界面如图 6 所示。在该界面内，高亮标注已标注的实体，点击头实体即可牵引出关系箭头，点击尾实体并选择恰当的关系标签即可完成关系的标注。关系标注的智能推荐将直接展示出推荐的标注结果，左键确认，右键取消。

当用户标注了一定量的文本，模型训练达到预先设置的停止条件时，工具不再显示智能推荐部分，而是直接对剩余文本进行预测，将标注结果直接展示在文本上，用户只需检查确认，若用户认为模型效果未达预期，也可对模型自动标注的结果进行更正。

结束语 本文基于实体与关系联合抽取的 UniRE 模型，设计并实现了一系列主动学习采样策略，提出了主动学习自动停止时机的确定方法，通过医疗领域 CBLUE 上的两个数据集验证了模型与方法的有效性。具体而言，使用基于 Entropy 的采样策略，在 CMeEE 数据集上使用 15%的数据量的训练效果即可达到全量数据训练效果的 95%，在更为复杂的 CMeIE 数据集上使用 33%的数据量可以达到全量数据训练效果的 90%，其减少标注成本的效果是显著的。而当模型在训练集上的准确率达到 98%时，使用基于训练集准确度的停止时机选取方法停止训练，其停止时机与采用验证集训练效果进行判断得到的停止时机较为接近。此外，本文设计并开发了一个智能文本标注系统，该系统具有友好易用的用户交互界面，使用智能推荐与主动学习最大限度地减少了用户标注的工作量，具有较强的实用性。

本文对知识抽取模型及主动学习算法的创新略有不足，未来研究可以从以下方面开展：首先，可以在模型中引入实体标签与关系标签的约束关系，用于规范模型训练，提升模型准确性；其次，可以考虑研发具有更好性能的主动学习算法。

参 考 文 献

[1] HANISCH D, FUNDEL K, MEVISSEN H T, et al. ProMiner: rule-based protein and gene entity recognition[J]. BMC Bioinformatics, 2005, 6(1): 1-9.

[2] ROCKTÄSCHEL T, WEIDLICH M, LESER U. ChemSpot: a hybrid system for chemical named entity recognition[J]. Bioinformatics, 2012, 28(12): 1633-1640.

[3] ZHENG S, WANG F, BAO H, et al. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 1227-1236.

[4] WEI Z, SU J, WANG Y, et al. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction[C]// Proceedings of the 58th Annual Meeting of the Association for Computational

Linguistics. 2020:1476-1488.
[5] WANG J, LU W. Two are Better than One: Joint Entity and Relation Extraction with Table-Sequence Encoders[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing(EMNLP). 2020:1706-1721.
[6] WANG Y, SUN C, WU Y, et al. UniRE: A Unified Label Space for Entity Relation Extraction[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing(Volume 1: Long Papers). 2021:220-231.
[7] SHEN Y, YUN H, LIPTON Z C, et al. Deep Active Learning for Named Entity Recognition[C]// Proceedings of the 2nd Workshop on Representation Learning for NLP. 2017:252-256.
[8] ZHDANOV F. Diverse mini-batch active learning [J]. arXiv: 1901.05954, 2019.
[9] ASH J T, ZHANG C, KRISHNAMURTHY A, et al. Deep batch active learning by diverse, uncertain gradient lower bounds [J]. arXiv: 1906.03671, 2019.
[10] ZHANG N, CHEN M, BI Z, et al. CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 7888-7915.
[11] HONGYING Z, WENXIN L, KUNLI Z, et al. Building a pediatric medical corpus: Word segmentation and named entity annotation[C]// 21st Workshop Chinese Lexical Semantics (CLSW 2020). Hong Kong, China, Revised Selected Papers 21. Springer International Publishing, 2021:652-664.
[12] GUAN T, ZAN H, ZHOU X, et al. CMeIE: Construction and evaluation of Chinese medical information extraction dataset [C]// 9th CCF International Conference Natural Language Processing and Chinese Computing(NLPCC 2020). 2020:270-282.

DING Hongxin, born in 2000, postgraduate. Her main research interests include knowledge graph, natural language processing and so on.

ZHAO Junfeng, born in 1974, Ph.D, research professor, is a member of China Computer Federation. Her main research interests include big data analysis, knowledge graph, urban computing and so on.

(责任编辑:喻藜)