数字人文视域下面向历史古籍的信息抽取方法研究

韩立帆[1,2]，季紫荆[1,2]，陈子睿[1,2]，王鑫[1,2]
1. 天津大学智能与计算学部，天津 300350；
2. 天津市认知计算与应用重点实验室，天津 300350

摘要
数字人文旨在采用现代计算机网络技术助力传统人文研究，文言历史古籍是进行历史研究和学习的重要基础，但由于其写作语言为文言文，与现代所用的白话文在语法和词义上均有较大差别，因此不易于阅读和理解。针对上述问题，提出基于预训练模型对历史古籍中的实体和关系等进行知识抽取的方法，从而有效获取历史古籍文本中蕴含的丰富信息。该模型首先采用多级预训练任务代替BERT原有的预训练任务，以充分捕获语义信息，此外在BERT模型的基础上添加了卷积层及句子级聚合等结构，以进一步优化生成的词表示。然后，针对文言文标注数据稀缺的问题，构建了一个面向历史古籍文本标注任务的众包系统，获取高质量、大规模的实体和关系数据，完成文言文知识抽取数据集的构建，评估模型性能，并对模型进行微调。在构建的数据集及GulianNER数据集上的实验证明了提出模型的有效性。

关键词
历史古籍；预训练模型；信息抽取；众包机制

中图分类号：TP391.1 文献标志码：A doi: 10.11959/j.issn.2096-0271.2022058

Research on information extraction methods for historical classics under the threshold of digital humanities

HAN Lifan[1,2], JI Zijing[1,2], CHEN Zirui[1,2], WANG Xin[1,2]
1. College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
2. Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin 300350, China

Abstract
Digital humanities aims to use modern computer network technology to help traditional humanities research. Classical Chinese historical books are the important basis for historical research and learning, but since their writing language is classical Chinese, it is quite different from the vernacular Chinese in grammar and meaning, so it is not easy to read and understand. In view of the above problems, the solution to extract entities and relations in historical books based on pre-trained models was proposed to obtain the rich information contained in historical texts effectively. The model

usedmulti-level pre-training tasks instead of BERT's original pre-training tasks to fully capture semantic information. And the model added some structures such as convolutional layers and sentence-level aggregations on the basis of the BERT model to optimize the generated word representation further. Then, in view of the scarcity of classical Chinese annotation data, a crowdsourcing system for the task of labeling historical classics was constructed, high-quality, large-scale entity and relation data was obtained and the classical Chinese knowledge extraction dataset was constructed. So it helped to evaluate the performance of the model and fine-tune the model. Experiments on the dataset constructed in this paper and on the GulianNER dataset demonstrated the effectiveness of the model proposed in this paper.

Key words
historical classics, pre-trained model, information extraction, crowdsourcing mechanism

0 引言

数字人文旨在将传统人文研究与现代计算机技术相结合，在我国，其研究多集中在图书情报领域，尤其是古籍的数字化研究方面。史书古籍不仅是文化的延续，还蕴含着丰富的信息，是历史研究和学习的重要基础，如“二十四史”中包含着海量的历史人物和事件，但由于其写作语言为文言文，与现代大众所使用的白话文区别较大，往往不易于阅读和理解。如果能准确高效地抽取出其中蕴含的实体和关系等知识信息，并形象展示，则能够有效推动人文领域的研究。

在众多知识表示方式中，知识图谱（knowledge graph，KG）作为一种语义网络，拥有极强的表达能力，可以灵活地对现实世界中的实体、概念、属性以及它们之间的关系进行建模。相比于其他结构知识库，知识图谱的构建及使用都更加接近人类的认知学习行为，因此对人类阅读更加友好。知识图谱旨在组织并可视化知识，其基础是命名实体识别（named entity recognition，NER）和关系提取（relation extraction，RE）这两项自然语言处理（natural language processing，NLP）任务。

近年来，自然语言处理技术的快速发展使人类使用自然语言与计算机进行通信成为可能。与此同时，深度学习（deep learning，DL）技术被广泛应用于各个领域，基于深度学习的预训练模型将自然语言处理带入一个新时代。预训练语言模型（pre-trained language model，PLM）极大地提升了语言信息表示的效果，成为目前自然语言处理领域的重要研究方向。预训练模型的目标在于使预训练好的模型处于良好的初始状态，在下游任务中具有更好的性能表现，同时减少训练开销，配合下游任务实现更快的收敛速度，从而有效提高模型性能，尤其是对一些训练数据比较稀缺的任务。

BERT（bidirectional encoder representations from transformer）[1]模型是预训练语言模型的代表之一，旨在通过联合调节上下文来预训练深度双向表示，主要分为两个阶段：预训练（pre-training）和微调（fine-tuning）。预训练阶段模型通过两种预训练任务来训练无标注数据，包括遮蔽语言模型（mask language model，MLM）任务和下一句话预测（next sentence predict，NSP）任务。模型在微调阶段使用预训练阶段的参数初始化，然后使用下游任务的标注数据来微调参数。由于BERT模型结构简单且有效性高，因此陆续出现了众多在其基础上进行

改进的模型，对于英语外的其他常用语言，研究人员也提出了针对不同语言的预训练模型。

针对中文的预训练语言模型研究近年来引起广泛关注，现有的中文预训练模型处理中文的能力已经在BERT模型的基础上得到进一步提升。然而，现有的中文预训练语言模型大多集中在白话文上，且现有的文言文预训练语言模型仅使用文言文语料进行预训练，没有针对性地修改模型结构和优化训练过程。因此，本文面向文言文特点构建了一个预训练语言模型，在BERT模型的基础上对预训练任务和模型结构进行优化，从而进一步提高预训练语言模型处理文言文的性能。

此外，目前现有的中文理解测评基准及数据集大多为白话文，无法针对性地微调模型使之适应文言文任务，同时无法准确评测模型处理文言文任务的性能。现有的文言文NER任务数据集来自第十九届中国计算语言学大会（the nineteenth China national conference on computational linguistics，CCL2020）“古联杯”古籍文献命名实体识别评测大赛，其标注数据仅包含“书名”及“其他专名”两类实体，且规模有限。因此，本文设计并构建了一个众包标注系统，结合群体智慧与领域知识实现标注的高效性和准确性，实现历史古籍文本中实体和关系的高精度抽取。根据系统获得的标注结果生成了文言文知识抽取数据集，包括建立在相应数据集上的细粒度NER任务和RE任务，数据集可用于微调当前自然语言处理主流的预训练语言模型，并评估模型，处理文言文的性能，同时能够为中国古代历史文献知识图谱构建提供数据支持。本文的整体技术框架如**图1**所示，在众包标注系统所得数据集上的实验证明了本文提出模型的有效性。

1 相关工作

1.1 数字人文视域下的文化遗产众包

数字人文是人文学科与计算机科学交叉研究衍生出的一个新领域，强调通过数字化重构的方式，以开放、共建和共享的形式将各类人文资源呈现于公众面前，近年来逐渐受到学术界和工业界的广泛关注，大量基于数字人文的文本挖掘、地理信息系统（geographic information system，GIS）、情感分析、可视化等应用开始出现。对文化遗产大数据的梳理离不开社会各界的共同努力，在数字化浪潮与文化建设需求的双重推动下，对文化遗产资源进行数字化、结构化、关联化等一系列运作，以开放数据的形式提供数字化服务，实现从静态资源保护向动态文化传承的转变，在保护和传播文化遗产的基础上让文化遗产资源得到有效利用。

众包一词最早由Howe J[2]提出，其核心含义是一家公司或机构将传统上由员工履行的职能以公开召集的形式外包给广泛而不确定的群体。早期的众包模式应用主要集中在商业领域。近年来，文化记忆机构逐渐意识到众包模式的价值，尝试引入众包模式开展一系列实践探索[3-4]。例如，利用大众力量进行各类文化遗产数据采集、标注或分类的工作。

从发起者角度来看，文化遗产众包项目可分为两大类：社会驱动型和组织驱动型。其中，社会驱动型项目数量不多且较少受到关注[5]；相较而言，组织驱动型项目更加广泛成熟。组织发起文化遗产的众包活动主要基于文化遗址和文化习俗的记录、保护与传承的需要。

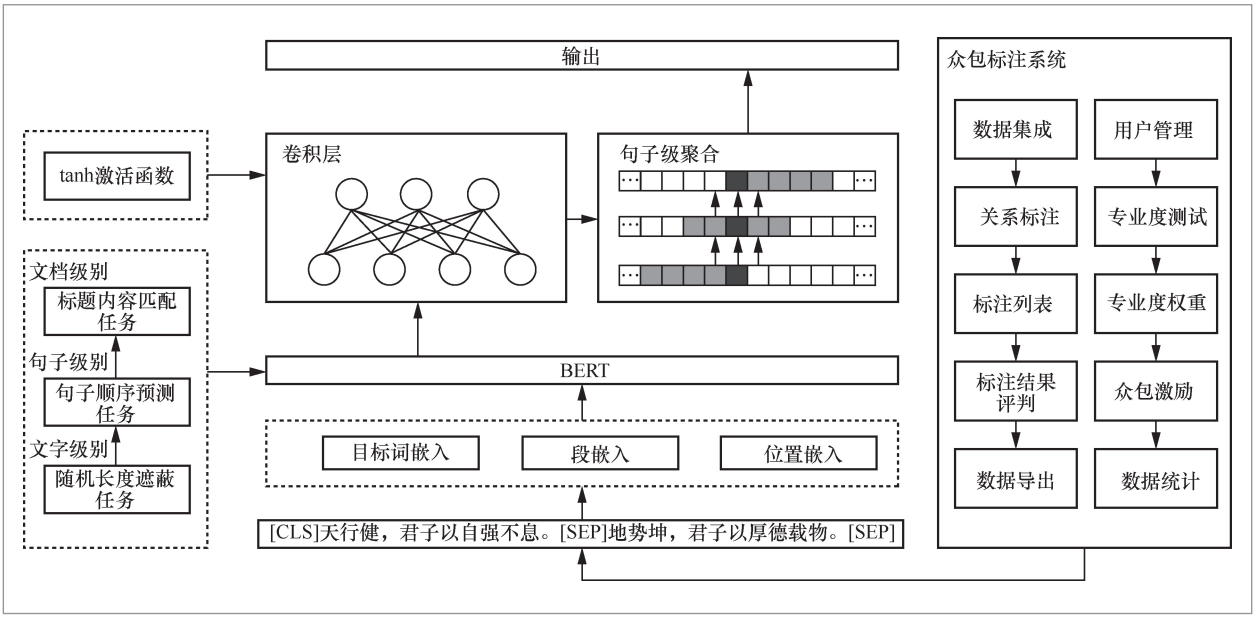

图1 整体技术框架

根据文化遗产资源类型的不同，组织驱动型文化遗产众包项目主要可细分为3类：馆藏资源建设型、文化遗址和景观保护型、非物质文化遗产保护型。其中，馆藏资源建设型文化遗产众包项目是指文化记忆机构在数字资源建设中引入众包模式，使公众深度参与这些特色资源的组织、交流和管理过程[6-7]。具体而言，众包模式在馆藏资源建设中的典型应用包括增加数据价值（如标记、评论）、提高数据质量（如数据更正）和补充数据内容（如创建并上传用户原创内容）。

然而，由于在数据化组织与管理、语义化分析、长期存储和开放获取等方面存在瓶颈[8-9]，文言文数字化转型之路困难重重，由一人或者一个小型团队独立完成的模式难以满足数据量大量增长的现实需求。而且经研究表明[10]，在共享文化环境和互动协作方式的众包模式下，非物质文化遗产的记录和保护获得了有力支持。鉴于此，本文通过设计构建一个众包标注系统，实现“二十四史”语料中实体和关系的高精度抽取，为文言文的保护与传承提供可行路径。

1.2 基于预训练模型的信息抽取

预训练模型是指预先设计好模型的网络结构，对输入数据进行编码训练，然后解码输出，提高模型的泛化能力。预训练完成后的模型可以根据下游任务的实际需要进行微调，相比从零开始训练模型节省了大量的开销。信息抽取是一种从文本数据中抽取特定信息的技术，主要包括两个子任务：命名实体识别和关系抽取。

早期的预训练模型技术基于传统的词向量嵌入[11-12]，2013年Mikolov T等人[13]提出的Word2vec模型对词嵌入向量进行了优化，包括了连续词袋模型（continuous bag-of-words model，CBOW）和跳字模型（continuous skip-gram model，Skip-gram）两种训练模式。相比传统词向量嵌入模型，Word2vec具有更低维度、更快运

算速度、更具通用性等优点，但同时也存在无法针对具体下游任务进行动态调整等缺点。

近年来，预训练模型占据信息抽取的主导地位，并取得最新成果。基于上下文的预训练模型开始流行，CoVe（contextualized word vectors）[14]模型可以从网络模型中获得上下文向量，并与词向量结合以提升模型性能。ELMo（embedding from language models）由Peters M等人[15]首次提出，并应用动态词向量方法实现复杂的语义表示，根据词所在语境上下文对词向量进行相应调整，动态生成词向量，解决了一词多义的问题。

Vaswani A等人[16]在ELMo模型和注意力（attention）机制的基础上提出了Transformer模型。该模型完全依赖于attention机制，没有使用诸如循环神经网络（recurrent neural network，RNN）和卷积神经网络（convolutional neural network，CNN）等较为流行的神经网络结构。attention机制一方面可以很好地处理较长序列，另一方面可以并行地处理数据。Transformer模型采用经典的编码器-解码器（encoder-decoder）结构，在编码器结构中将输入文本序列处理为一个连续的输出序列，在解码器结构中将这个输出序列进行处理，并输出结果。

Transformer模型在诸多任务中的应用效果表明，它的特征提取能力强于长短期记忆(long short-time memory，LSTM)[17]神经网络结构，因此Radford A等人[18]基于Transformer模型提出了GPT（generative pre-training）模型，该模型利用了Transformer模型中的Decoder结构，并且训练阶段与ELMo相同，均采用两阶段训练模式，先通过大量语料完成模型预训练，再针对具体下游任务完成第二阶段的微调训练。

ELMo和GPT模型均为自回归模型，ELMo虽然使用了双向长短时记忆（bidirectional long short-term memory，Bi-LSTM）[19]网络获取双向语义信息，但其将方向相反的两个网络进行叠加的做法不能真正实现对文本的双向语义理解。谷歌（Google）在2018年提出了BERT模型，该模型基于Transformer模型的Encoder结构，使用完全双向的语言模型结构，同样采用预训练和下游任务两阶段的训练模式。BERT模型的出现大幅提升了自然语言处理任务的效果。基于对BERT模型的优化，RoBERTa[20]模型将BERT模型中的静态掩码替换成了动态掩码，即对每次输入的句子进行随机掩码，并移除BERT模型中的下一句话预测任务，进一步增强了模型在文本推理任务中的表现。同时，基于BERT模型衍生的许多预训练模型[21-22]，为自然语言处理任务中大量缺乏标注数据的任务提供了新的思路。2019年 Cui Y M等人[23]提出的BERT-wwm模型将全词掩码（whole word masking，WWM）的方法应用至中文预训练模型，取得了更优的实验效果。

王东波等人[24]以《四库全书》为训练集构建的SikuBERT和SikuRoBERTa预训练语言模型在古文NLP任务上表现出了优秀的学习与泛化能力。但是，目前基于预训练模型完成文言文信息抽取的研究工作仍然还有很大的提升空间。本文基于BERT预训练语言模型构建了一个面向文言文语义特点的预训练语言模型，通过优化BERT模型的预训练任务和模型结构，提高预训练语言模型处理文言文的性能，并在此基础上进一步完成文言文语言理解数据集上细粒度的NER任务和RE任务。

2 基于预训练模型的知识抽取

2.1 预训练任务

BERT模型采用遮蔽语言模型和下一句话预测这两个任务对模型进行预训练。具体来说，遮蔽语言模型任务对输入文本中15%的字进行随机遮蔽，遮蔽部分以80%的概率将其改变为“[MASK]”标签，以10%的概率将其替换为随机字，以10%的概率保持不变，之后让模型对遮蔽的内容进行预测。下一句话预测任务则从语料库中抽取一个语句，再以50%的概率抽取它之后紧接着的语句，以50%的概率随机抽取一个其他的语句，让模型判断这两个语句是否是相邻语句。两个任务分别学习输入文本的词级别信息和句子级别信息，目前已被证明均有提升空间。遮蔽语言模型采用类似完形填空的方式让模型学习预测缺失字，但没有考虑到词语边界信息；而下一句话预测任务难度较小，由于抽取到的两个语句很可能并不属于同一话题，因此较容易识别其是否衔接，不利于模型学习句子之间的联系。

本文针对文言文语料的特点对原始的预训练任务进行了优化，分别采用词级别随机长度遮蔽任务、句子级别句子顺序预测任务以及文档级别标题内容匹配任务以充分捕获多级语义。具体来说，文言文中单字往往可以表达完整含义，无须对其进行分词，因此本文采用一种已被证明简单有效的随机长度遮蔽任务，并随机选择长度为1到最大长度的目标进行遮蔽。如果将最大长度定义为N，则遮蔽片段长度为1~N，此时遮蔽片段长度为n的概率如式（1）所示：

$$p(n)=\frac{1/n}{\sum_{k=1}^{N}1/k} \tag{1}$$

其中，n和k的取值范围均为1~N。在本文中，最大长度N为3。

此外，本文使用句子顺序预测任务代替BERT模型的下一句话预测任务。该任务将来自同一文档的两个连续文段作为正例，以50%的概率将两个连续段落的顺序交换作为反例，避免文段主题的差别，促使模型专注于学习句子间的连贯性。

最后，为了学习到更高级别的语义信息，本文提出文档级别标题内容匹配任务。具体来说，考虑到古代诗词往往篇幅较短、标题通常包含诗词主题的特点，该任务将中国古代诗词数据集作为训练语料。该任务是一种类似于句子顺序预测任务的二元分类任务，用于捕获高级语义信息。具体来说，该任务将标题和内容匹配的诗词作为正例，将50%的概率打乱诗词的标题与内容之间的匹配作为反例，使模型学习标题与内容的语义关联，捕获更高级别的语义信息。

2.2 模型结构

首先，模型对于输入语料中的每个文字生成3个部分词嵌入，即目标词嵌入、段嵌入以及位置嵌入，叠加后输入类似于BERT模型结构的Transfomer编码器进行处理。

对于输出的词向量，为了获得更多可学习的表示，本文引入一个卷积层，将预训练语言模型生成的语料表示输入该卷积层，使用激活函数非线性地将词嵌入转换为更高级别的特征。对于字符向量$\boldsymbol{x}_t$，经过卷积层生成的词嵌入定义如式（2）所示：

$$\boldsymbol{x}_t=\tanh(\boldsymbol{W}_t\boldsymbol{x}_t+\boldsymbol{b}_t) \tag{2}$$

其中，W_t表示权重矩阵，b_t表示偏置向量。

此外，为了进一步增强词表示，本文利用滑动窗口机制，设计了句子级聚合，以有效地获取相邻字符信息。具体来说，本文人工设置窗口大小，窗口在目标句划定的范围内滑动，从窗口第一次包含目标字符开始，到窗口最后一次包含目标字符结束，所有经过的词及目标字符本身都被视为目标字符的邻居。在滑动过程中需要考虑两种特殊情况，即如果目标词是句子中的第一个或最后一个词，则窗口滑动范围等于窗口大小。出于简洁性考虑，在实验中使用平均聚合方法，在给定窗口中聚合词向量的邻居信息。本文将聚合过程定义为AGG函数，则字符向量$\boldsymbol{x}_t$在窗口尺寸w下的聚合结果$\boldsymbol{h}_t^w$定义如式（3）所示：

$$\begin{aligned}\boldsymbol{h}_t^w &= \mathrm{AGG}\\ &(\{x_j \mid p_j \in [\max(1, p_t - w + 1),\\ &\min(s, p_t + w - 1)]\})\end{aligned} \tag{3}$$

其中，$\boldsymbol{x}_j$是邻居字符向量，p_t和p_j分别表示字符向量和邻居字符向量的位置，s表示句子的长度。由线性变换得到$\boldsymbol{x}_t$的新表示$\boldsymbol{h}_t$定义如式（4）所示：

$$\boldsymbol{h}_t = \sigma\left(\boldsymbol{W}_s \times \boldsymbol{h}_t^w\right) \tag{4}$$

其中，$\boldsymbol{W}_s$是一个可学习的权重矩阵，σ是一个激活函数，如ReLU函数。

3 基于众包系统的知识抽取数据集构建

3.1 众包系统设计与实现

本文针对历史古籍标注任务专业性较强的特点，设计并构建了一个众包标注系统，引入“二十四史”的全部文本，允许工作者标注其中的实体和关系。不同于现有的众包系统，由于该标注任务需要工作者具备领域知识，因此本文将工作者专业度引入系统，以得到更准确的标注结果。具体来说，工作者初次登录系统时，系统需要对其进行专业度判断，同时在答案整合和众包激励分配的阶段均将专业度纳入考虑。此外，目前的众包系统大多注重任务的分配，系统中的标注任务多以题目的形式呈现，并尽可能通过任务分配算法交给能够准确作答的工作者。而本文的系统中，标注任务以文本的形式呈现，并向每名工作者开放相同任务，即“二十四史”的全部内容均在系统中呈现，工作者可以自行选择感兴趣的章节，也可以对同一文本进行不同的标注，最大限度地发挥群体智慧。

众包系统的工作者标注界面如**图2**所示，每位工作者可以从左侧的树形目录中选择感兴趣的篇章进行标注，系统支持实体和关系的标注，并将标注出的实体用带有背景色的方框显示，标注出的关系用斜体并加下划线显示。每一页对应历史古籍文本中的一个段落，在每个结束标点处换行，方便工作者进行阅读和定位。

由于本系统涉及的标注任务具有较强的专业性，需要在工作者初次登录系统时就对其专业能力进行判断，以了解该工作者是否能够胜任本系统开放的标注任务。因此，本系统引入了大多现有众包系统未纳入考虑的工作者专业度，并定义了两种工作者类型，即“专家工作者”和“普通工作者”，同时定义了两种判断方法。

对于已知的专业度较高的工作者，如高校的教师、学生等，在将其信息录入数据库时，可以直接将其类型定义为“专家工作者”。而对于未知工作者，如社会上的历史爱好者等，系统准备了具有标准答案的测试题目，要求工作者首次登录系统时进行作答，根据工作者的答题准确率和题

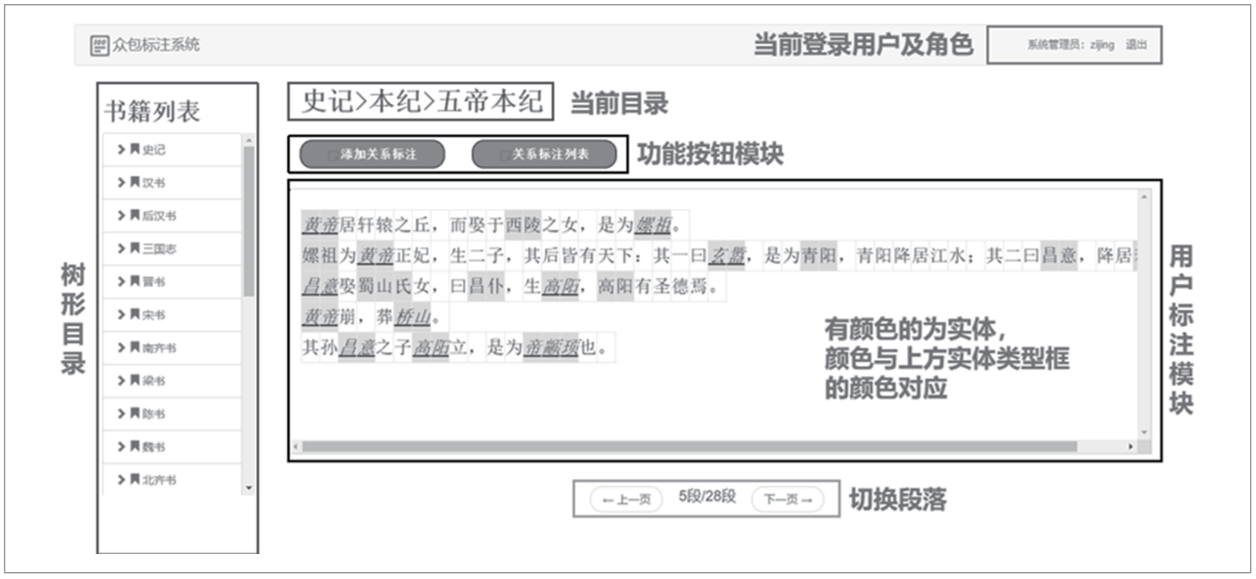

图 2 众包系统的工作者标注页面

目难度综合计算该工作者的专业度，具体计算方式如下。①选取若干志愿者（在本文中均为历史或计算机专业的学生），准备若干具有标准答案的题目请志愿者进行作答，根据其答题情况为每道题目赋难度初始值，难度定义为答错的志愿者数量与参与答题志愿者总数的比值（取值范围为[0,1]）。②工作者进入系统后，题目难度动态变化，计算方式仍然为答错的工作者总数与参与答题工作者总数的比值，此时的工作者总数为注册工作者的数量和志愿者数量的总和。③题目的分数与难度值成正比，定义为难度值乘10并向上取整，将所有题目分数之和定义为总分。如果工作者得到的分数能够高于总分的60%，将其定义为专家工作者，反之，则将其定义为普通工作者。

众包系统中的专业度测试页面如**图3**所示，工作者首次登录系统时将跳转到该页面进行专业度测试，完成页面的题目后点击“提交”按钮，系统返回工作者得分占总分的百分比及工作者类型。

对于同一题目，若工作者具有不同的回答，现有系统往往采用投票策略，以多数工作者的作答为最终结果。后续也有很多研究工作将工作者在系统中完成标注任务的准确率纳入考虑，以获取更加准确的结果，而对于本文系统中定义的历史古籍中的实体和关系标注任务，专业度高的工作者更有可能做出正确的标注。因此，不同于现有方法，为了确保结果的准确性，本系统在答案整合阶段充分考虑了工作者专业度。

具体来说，系统允许工作者修改页面上的现有标注，并在工作者进行标注时将工作者ID、标注时间以及标注内容等信息均录入数据库。如果多名工作者对同一个实体或实体对有不同的标注，在页面上会展示最新的标注结果，而在数据库中将分别保存它们，即出现新的标注并不会覆盖之前的标注。在下载数据时，若对应同一文本存在多条标注记录，即同一文本存在不同的标注内容，则将进行基于工作者专业度的答案整合。具体来说，系统为专家工作者赋予双倍于普通工作者的权重，并采用加权多数投票策略来获得最终结果，而

图 3 众包系统中的专业度测试页面

特别的是，若工作者将页面上的现有标注删除，系统同样会将该操作录入数据库，并认为此工作者对该文本的判断为非实体或实体对不存在关系。

大多现有众包系统在计算工作者的奖励时考虑了其标注数量及准确率，而本文在该基础上，将工作者专业度纳入考虑，提出了一种新的众包激励机制，并以固定的周期结算奖励。具体来说，简单地将答案整合后的最终结果视为正确结果，如果工作者的标注与正确结果相同，则给予奖励，否则不给予奖励。此外，认为专家工作者做出正确标注的可能性更高，因此为了激励其积极进行标注，给予其双倍于普通工作者的奖励。最后，为了调动工作者尽可能准确地完成更多的标注，该系统还对标注的数量和正确率设置了阈值，对超过该阈值的用户给予多倍奖励。

例如，若将一次实体标注或关系标注的单价设为p，标注数量阈值设为a_t，标注准确率阈值设为c_t，这时如果一名普通工作者在某一奖励分配周期内完成了n个标注，其中正确标注m个，且该工作者的标注数量与标注准确率均超过了系统设定的阈值，则该工作者将获得的奖励reward计算方式如式（5）所示：

$$\text{reward} = m \times \left(1 + \frac{m}{n} - c_t\right) \times \frac{n}{a_t} \times p \qquad (5)$$

3.2 知识抽取数据集构建

基于众包标注系统的实体和关系标注结果，构建了一个由NER和RE任务及其相应数据集组成的文言文知识抽取数据集。细粒度NER任务数据集由文本文件和标签文件组成。文本文件与标签文件逐行对应，共定义6类实体：人名（RER）、地名（LOC）、职位名（POS）、组织名（ORG）、书名和战争名。在标签文件中，采用BIO标注法对文本进行标注，对标注为实体首字的文本赋予“B-”标签，对标注为实体中间字或尾字的文本赋予“I-”标签，对非实体的文字赋予“O”标签。NER任务数据集的统计信息见**表1**。

RE任务数据集的统计信息见**表2**，共包括7类关系：组织名-组织名、地名-

表1 NER任务数据集的统计信息

实体类型	训练集/个	校验集/个	测试集/个	总数/个
人名	9 467	1 267	701	11 435
地名	2 962	391	167	3 520
职位名	1 750	242	139	2 131
组织名	1 698	266	100	2 064
其他	110	18	9	137

表2 RE任务数据集的统计信息

关系类型	训练集/个	校验集/个	测试集/个	总数/个
人名-人名	1 139	324	130	1 593
人名-地名	462	129	53	644
人名-职位名	1 093	319	162	1 574
人名-组织名	231	60	38	329
其他	157	40	28	225

组织名、人名-人名（PER-PER）、人名-地名（PER-LOC）、人名-组织名（PER-ORG）、人名-职位名（PER-POS）和地名-地名。基于原始数据集，本文可以生成一个由句子和关系文件组成的关系分类数据集，该数据集中，句子文件和关系文件逐行对应，表示每一个句子及其所包含的关系。此外还可以生成一个类似于NER任务数据集的序列标记数据集，该数据集同样由文本文件和标签文件组成，但这时，生成的标签不再是实体类别标签，而是标志其是某关系的主体或客体的标签。

4 实验及结果分析

4.1 参数设置及评价指标

在微调阶段，除批量大小（batch size）、学习率（learning rate）和训练轮数（epoch）外，其他超参数均与BERT预训练阶段所使用的超参数相同。实验结果表明，能够在微调阶段获得较好效果的超参数取值如下：batch size取32；learning rate取5×10^{-5}、3×10^{-5}、2×10^{-5}；epoch取3~10。

本文在实验中将F1值作为衡量模型性能表现的评价指标，它综合考虑了精确率和召回率。如果模型能够在测试集上取得较好的性能，可以考虑使用模型自动抽取未标注文本中的实体和关系，以进一步扩展数据集；否则，迭代从系统中获取新标注的实体和关系再对模型进行微调，直到模型能够在文言文任务上取得出色表现。

4.2 数据集

本文除了采用由众包系统中获取的数据构建的数据集外（介绍详见第3.2节），还采用了CCL2020“古联杯”古籍文献命名实体识别评测大赛主办方提供的GulianNER数据集，该数据集定义了书名（BOOK）和其他专名（OTHER）两类实体，数据集的统计信息见**表3**。

表 3 GulianNER 数据集的统计信息

实体类型	训练集/个	校验集/个	测试集/个	总数/个
书名	27 445	11 531	5 633	44 609
其他专名	91 917	20 972	5 552	118 441

4.3 实验结果与实验分析

本文在基准测试中评估了以下预训练模型：BERT-Base、BERT-wwm、RoBERTa-zh和Zhongkeyuan-BERT（以下简称ZKY-BERT），简要介绍如下。

- BERT-Base：谷歌人工智能研究院于2018年10月提出的预训练模型，是NLP发展史上具有里程碑意义的模型成果。
- BERT-wwm：采用全词遮蔽任务，引入词边界信息，由遮蔽随机译字（token）改为分词后对完整的词进行遮蔽。
- RoBERTa_zh：使用更大的模型参数，更大的bacth size和更多的训练数据。此外，在训练方法中，去除了下一句预测任务，采用了动态遮蔽方法，加强了训练实例的随机性。
- ZKY-BERT：使用殆知阁语料和唐诗宋词数据集等文言文语料进行进一步的预训练，将最大句子长度从128修改为512。另外，设立了受限波束搜索以排除非法转换。

在6类实体数据集上的实验结果如**图4**所示。可以观察到，在处理细粒度NER时，本文模型能够取得最好的性能表现，在文言文语料库上训练的ZKY-BERT模型表现和适应中文特点的BERT-wwm模型也能取得较好性能，模型之间的性能表现差距较大。

由于战争名和书名两类实体数量较少，为了进一步提升模型的性能，本文采用了去除这两类实体的数据集进行实验，结果如**图5**所示。可以观察到，由于实体类型减少，预训练模型均表现出了相对较好的性能，且模型之间的性能差距缩小。

在GulianNER数据集上的实验结果如**图6**所示，由于该数据集中包含的实体类型较少且数据规模较大，模型均能取得较好的性能表现。可以观察到，本文模型依然能取得最佳性能，在文言文上训练过的ZKY-BERT模型次之，模型之间的性能差距缩小。

对于RE任务，本文将其拆分为两个子任务：关系分类和序列标记。实验表明，基线模型在关系分类任务上可以达到47.61%的准确率，而由于关系类型较多且数据较为分散，在序列标注任务上各模型都不能取得较好的性能表现。

5 结束语

为了基于预训练模型实现历史古籍中实体和关系数据的抽取，助力传统人文研究，并为知识图谱的构建提供数据基础，本文提出基于BERT模型对其预训练任务和模型结构均进行优化的方法。针对文言文知识抽取任务的特点设计多级预训练任务，并添加卷积层及句子级聚合等结构进一步优化词表示。同时，构建了一个基于工作者专业度的众包标注系统，以实现对古籍文本中实体和关系的标注，从而构建一个文言文上的语言理解测评基准，对模型的性能进行评估和微调。实验证明了本文提出的模型相较于其他基线模型在处理文言文任务的性能上有所提高。

由于基准集数据量较小，本文的模型在知识抽取任务上的性能表现仍有较大提升空间。在未来工作中，笔者将探索如何高效获取更多标注数据，并进一步探索如何提升模型在文言文上的性能表现，以推进传统人文领域的研究。

参考文献:

[1] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of NAACL-HLT. [S.l.:s.n.], 2019: 4171-4186.

[2] HOWE J. The rise of crowdsourcing[J]. Wired, 2006, 14(6): 176-183.

[3] HOLLEY R. Crowdsourcing: how and why should libraries do it?[J]. D-Lib Magazine, 2010, 16(3/4): 1-21.

[4] OOMEN J, AROYO L. Crowdsourcing in the cultural heritage domain: opportunities and challenges[C]// Proceedings of the 5th International Conference on Communities and Technologies. [S.l.:s.n.], 2011: 138-149.

[5] TERRAS M. Digital curiosities: resource creation via amateur digitization[J]. Literary and Linguistic Computing, 2010, 25(4): 425-438.

[6] RIDGE M. Citizen history and its discontents[C]//Proceedings of 2014 IHR Digital History Seminar. [S.l.:s.n.], 2014: 1-13.

[7] ZHANG X H, SONG S J, ZHAO Y C, et al. Motivations of volunteers in the Transcribe Sheng project: a grounded theory approach[J]. Proceedings of the Association for Information Science and Technology, 2018, 55(1): 951-953.

[8] RIDGE M. From tagging to theorizing: deepening engagement with cultural heritage through crowdsourcing[J]. Curator: the Museum Journal, 2013, 56(4): 435-450.

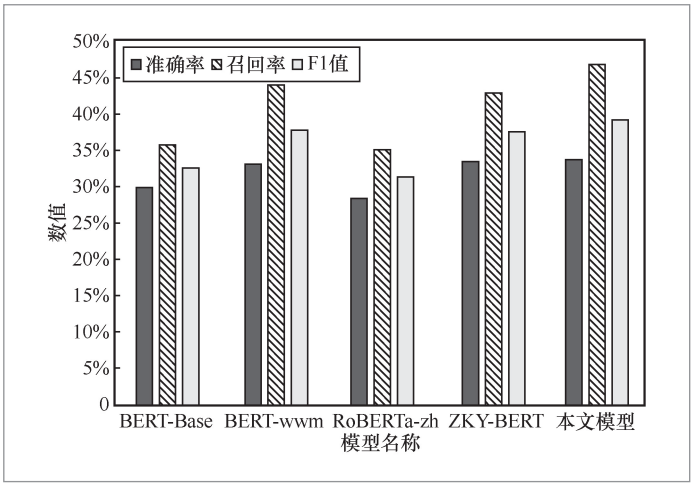

图 4 在 6 类实体数据集上的实验结果

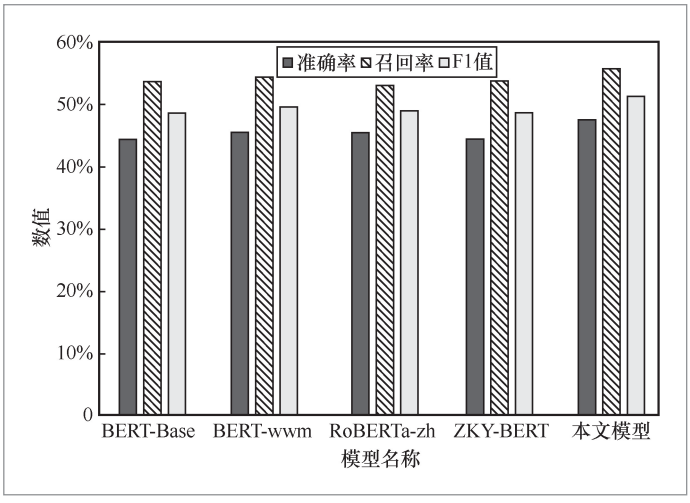

图 5 在 4 类实体数据集上的实验结果

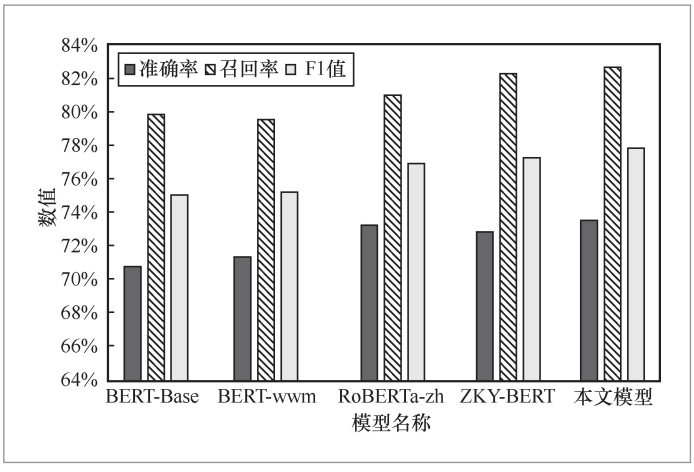

图 6 在 GulianNER 数据集上的实验结果

[9] DANIELS C, HOLTZE T L, HOWARD R I, et al. Community as resource: crowdsourcing transcription of an historic newspaper[J]. Journal of Electronic Resources Librarianship, 2014, 26(1): 36–48.

[10] CONCILIO G, VITELLIO I. Co–creating intangible cultural heritage by crowd–mapping: the case of mappi[na][C]//Proceedings of 2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a Better Tomorrow. Piscataway: IEEE Press, 2016: 1–5.

[11] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back–propagating errors[J]. Nature, 1986, 323(6088): 533–536.

[12] HINTON G E,MCCLELLAND J L, RUMELHART D E. Distributed representations[M]. Cambridge: MIT Press, 1986: 77–109.

[13] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint, 2013, arXiv:1301.3781.

[14] MCCANN B, BRADBURY J, XIONG C, et al. Learned in translation: contextualized word vectors[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2017: 6297–6308.

[15] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Stroudsburg: Association for Computational Linguistics, 2018: 2227–2237.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30.

[17] STAUDEMEYER R C, MORRIS E R. Understanding LSTM–a tutorial into long short–term memory recurrent neural networks[J]. arXiv preprint, 2019, arXiv: 1909.09586.

[18] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Imporoving language understanding by generative pre–training[Z]. 2018.

[19] MA X Z, HOVY E. End–to–end sequence labeling via Bi–directional LSTM–CNNs–CRF[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2016.

[20] LIU Y, OTT M, GOYAL N, et al. RoBERta: a robustly optimized bert pretraining approach[J]. arXiv preprint, 2019, arXiv: 1907.11692.

[21] LAN Z, CHEN M, GOODMAN S, et al. Albert: a lite bert for self–supervised learning of language representations[J]. arXiv preprint, 2019, arXiv: 1909.11942.

[22] YANG Z, DAI Z, YANG Y, et al. Xlnet: generalized autoregressive pretraining for language understanding[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2019: 5753–5763.

[23] CUI Y M, CHE W X, LIU T, et al. Pre–training with whole word masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504–3514.

[24] 王东波, 刘畅, 朱子赫, 等. SikuBERT与SikuRoBERTa：面向数字人文的《四库全书》预训练模型构建及应用研究[J]. 图书馆论坛, 2022, 42(6): 31–43.
WANG D B, LIU C, ZHU Z H, et al. Construction and application of pre–trained models of Siku Quanshu in orientation to digital humanities[J]. Library Tribune, 2022, 42(6): 31–43.

作者简介

韩立帆（1999– ），男，天津大学智能与计算学部硕士生，主要研究方向为自然语言处理、知识图谱构建。

季紫荆（1997– ），女，天津大学智能与计算学部硕士生，主要研究方向为自然语言处理、知识图谱构建。

陈子睿（1998– ），男，天津大学智能与计算学部硕士生，主要研究方向为知识表示学习、知识图谱问答、知识图谱构建。

王鑫（1981– ），男，博士，天津大学智能与计算学部教授、博士生导师，主要研究方向为知识图谱数据管理、图数据库、大规模知识处理。

收稿日期： 2022-03-13

通信作者： 王鑫，wangx@tju.edu.cn

基金项目： 科技创新2030—“新一代人工智能”重大项目 (No.2020AAA0108504)；国家自然科学基金资助项目 (No.61972275)

Foundation Items: Science and Technology Innovation 2030 “New Generation Artificial Intelligence” Major Project (No.2020AAA0108504), The National Natural Science Foundation of China (No.61972275)