融入篇章信息的文学作品命名实体识别

贾玉祥[1]，晁睿[1]，眷红英[1]，窦华溢[1]，曹帅[1,2]，徐硕[1,2]
1.郑州大学，河南　郑州
2.郑州中业科技股份有限公司，河南　郑州
ieyxjia@zzu.edu.cn；zzuruichao@163.com；iehyzan@zzu.edu.cn；
1014728581@qq.com；iscshuai@163.com；125645650@qq.com

摘要

命名实体识别是文学作品智能分析的基础性工作，当前文学领域命名实体识别的研究还较薄弱，一个主要的原因是缺乏标注语料。本文从金庸小说入手，对两部小说180余万字进行了命名实体的标注，共标注4类实体5万多个。针对小说文本的特点，本文提出融入篇章信息的命名实体识别模型，引入篇章字典保存汉字的历史状态，利用可信度计算融合BiGRU-CRF与Transformer模型。实验结果表明，利用篇章信息有效地提升了命名实体识别的效果。最后，我们还探讨了命名实体识别在小说社会网络构建中的应用。

关键词： 文学作品；命名实体识别；篇章信息

Document-level Literary Named Entity Recognition

Yuxiang Jia[1],Rui Chao[1],Hongying Zan[1],Huayi Dou[1],Shuai Cao[1,2],Shuo Xu[1,2]
1.Zhengzhou University, Zhengzhou, Henan, China
2.Zhengzhou Zoneyet Technology Co., Ltd., Zhengzhou, Henan, China
ieyxjia@zzu.edu.cn；zzuruichao@163.com；iehyzan@zzu.edu.cn；
1014728581@qq.com；iscshuai@163.com；125645650@qq.com

Abstract

Named entity recognition is the basic building block of intelligent analysis of literary works. At present, the research of named entity recognition in literary field is relatively backward. One of the main reasons is the lack of annotated datasets.We annotate over 50 thousands named entities of four types from about 1.8 million words of two Jin Yong's novels. According to the characteristics of novel text, this paper proposes a document-level named entity recognition model , using a document-level dictionary to save the historical state of Chinese characters, and uses credibility calculation to fuse BiGRU-CRF and Transformer model. The experimental results show that the use of document information can effectively improve the performance of named entity recognition. Finally, we also explore the application of named entity recognition in the construction of novel social network.

Keywords: Literary text , Named entity recognition , Document level

1 引言

命名实体识别 (li et al., 2003)是信息抽取的基础任务之一，其主要目标是将文本中的人名、地名、机构名等实体名识别出来，以服务于自动问答、机器翻译、文本分析等下游任务。使用命名实体识别技术，识别出文学作品中人物等实体，对文学作品的智能分析和数据挖掘具有重要意义，如人物建模 (Bamman et al., 2013)、社会网络构建 (Labatut and Xavier, 2019)、事件抽取 (Matthew et al., 2019)等任务。文学领域的命名实体识别，目前还是一个具有挑战性的任务 (Jockers, 2013)。主要原因有：文学作品语料与通用领域语料的差异大，不同作者写作风格不同，模型难以泛化；作品中人物等实体复杂多变，模型难以学习特征；文学领域的命名实体识别研究较少，缺乏大规模高质量的标注语料。

通用的命名实体识别方法大多是针对句子级的数据，这些方法能够捕捉到句子内部的依赖关系，但是忽略了句间的长距离依赖，而篇幅长正是小说等文学作品的一大特点。文学作品中的人物、地点等实体通常不是孤立存在的，同一篇章中的实体相互交织形成一个整体的网络 (Lin et al., 2018)。以往的命名实体识别方法只关注实体本身，忽略了篇章中实体之间的联系。本文模型将篇章中的实体进行一致性把控和识别，能够有效提升文学作品命名实体识别的效果。

本文的主要研究贡献包括以下两个方面：

（1）构建了基于两部金庸武侠小说的命名实体识别数据集，并将开放共享。

（2）针对文学作品特性，提出一种融入篇章信息的命名实体识别模型，实验表明该方法能够有效提升文学作品命名实体识别的效果。

2 相关工作

近些年深度学习模型在自然语言处理领域取得了广泛的应用和良好的效果 (Thomas and Sangeetha, 2020)。Collobert等人(2011)首次提出基于深度学习神经网络的命名实体识别模型，该模型能够在大规模的真实数据上获得较好的命名实体识别效果，但该模型无法捕捉长距离的依赖关系。Kuru等人(2016)使用LSTM提取全局特征，在多种不同语言上均取得了良好命名实体识别效果。柏冰等人(2018)将双向LSTM模型和CRF模型进行结合，在1998年人民日报语料库上取得了较好结果。Zhang等人(2018)在序列标注模型基础上融入潜在的词汇信息，提出了Lattice LSTM模型。Liu等人(2018)提出一种LM-LSTM-CRF任务感知型模型，在CoNLL-03命名实体识别数据集上F1值达到了91.71%。王月等人(2019)提出一种BERT-BiLSTM-Attention-CRF模型，在警情文本命名实体识别上准确率达91%。陈茹等人(2020)充分考虑到文本层次化结构对实体识别的重要性，提出了IDC-HSAN模型。Gui等人(2020)通过对篇章级数据中的标签进行一致性把控，提出一种篇章级命名实体识别方法，在多个英文数据集上达到最优结果。

以往命名实体识别研究大都聚焦在通用领域语料，针对文学作品领域的尝试较少。Vala等人(2015)提出一种基于图的pipline模型来识别文学作品中的人物，在多个数据集上获得很好的结果。同时提出一个衡量人物识别性能的评价框架，为以后相关工作提供了评价标准。Brooke等人(2016)提出一种LitNER模型用于小说命名实体识别，该模型基于bootstrap方法，使用无监督方式进行训练，实验表明在给定文本全部上下文的条件下，该模型优于有监督方法。Xu等人(2017)为解决中文文学文本命名实体识别数据匮乏的问题，基于700余篇文学领域的文章构建了一个面向中文文学作品的命名实体识别和关系抽取数据集。谢韬(2018)通过对宋词和史记等古文语料进行研究，首先基于Apriori算法和LSTM模型对语料进行分词处理，再通过LSTM和CRF模型对数据进行命名实体识别，实验表明，该方法能够有效识别古文中的重要实体。Bamman等人(2019)依照ACE标注规范，对100部英文小说进行命名实体标注，标注范围为每一部小说的前1000个词，在该数据集上训练能够有效提升文学领域命名实体识别的效果。

3 中文小说NER数据集

3.1 人物/PER

人物主要指具体角色的人名，是文学作品中的核心实体。如表1所示，“郭靖心想不错”中包含的人物“郭靖”，“丘处机和她在终南山上比邻而居”中的“丘处机”均为人物实体。需要注意的是文学作品中通常包含非主要人物的集合、亲属称谓、人名指代，这些不在人物实体的标注范

围内。例如：“两株大松树下围着一堆村民”中的“村民”，“却不提父亲已自刎身死之事”中的“父亲”，均不做标注。

小说文本	人物实体
郭靖心想不错	郭靖
丘处机和她在终南山上比邻而居	丘处机

表 1. 人物实体

3.2 地点/LOC

地点主要指故事情节发生的地点或者对话中提及的地点，地点实体包括的范围较广。如表2所示，“从西域带来了这八盆兰花”中的“西域”为范围较大的地点实体，“临安府牛家村村民郭啸天、杨铁心二犯”中的“临安府”、“牛家村”为范围较小的地点实体，需注意“临安府”、“牛家村”应分开标注。

小说文本	地点实体
从**西域**带来了这八盆兰花	西域
临安府牛家村村民郭啸天、杨铁心二犯	临安府、牛家村

表 2. 地点实体

3.3 组织/ORG

组织主要指门派或者帮会，主要特点：成员们的武功都是一脉相传的，常以“派”，“门”，“教”，“帮”，“会”结尾。如表3所示，“要讲武功，终究全真教是正宗”中的“全真教”，“丐帮却号称江湖上第一大帮”中的“丐帮”均为组织实体。

小说文本	组织实体
要讲武功，终究**全真教**是正宗	全真教
丐帮却号称江湖上第一大帮	丐帮

表 3. 组织实体

3.4 武器/WEP

武器主要包括兵器和武功。兵器指的是某人所使用的特定武器，如表4所示，“丐帮中规矩，见了打狗棒如见帮主本人”中的“打狗棒”为武器中的兵器实体。需注意单字的兵器实体不做标注。武功指的是某个人所使用的特定武功，如表4所示，“贫道就以太极拳中的招数和他拆几手”中的“太极拳”为武器中的武功实体。

小说文本	武器实体
丐帮中规矩，见了**打狗棒**如见帮主本人	打狗棒
贫道就以**太极拳**中的招数和他拆几手	太极拳

表 4. 武器实体

3.5 数据集统计

数据集总体统计如表5所示，人物实体占比最大。高频人物实体如图1所示，可以看出两部小说的主要人物“郭靖”“杨过”“黄蓉”“小龙女”出现频次最高，符合小说的主要人物特性。高频地点实体如图2所示，其中“蒙古”和“襄阳”为主要故事情节发生地点。高频组织实体如图3所示，其中“桃花岛”为组织实体，而并非一个地点，“全真教”、“全真派”和“全真”为同一组织的不同称呼，后续将进行共指消解的标注。高频武器实体如图4，其中“长剑”为多数人使用兵器，故出现频次最多，“九阴真经”为两部小说中重要武功，出现频次为第二。

实体类型	实体数量	比例	去重后数量
人物	44116	84.3%	942
地点	2409	4.6%	284
组织	1500	2.9%	39
武器	4330	8.3%	702

表 5. 数据集总体统计

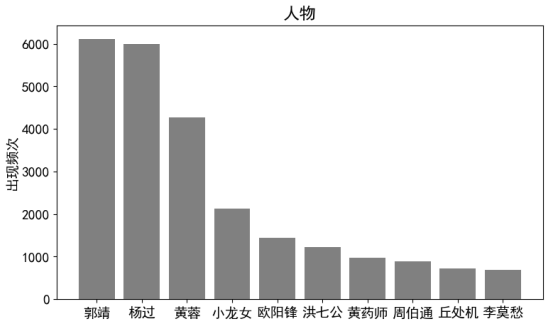

图 1. 高频人物实体

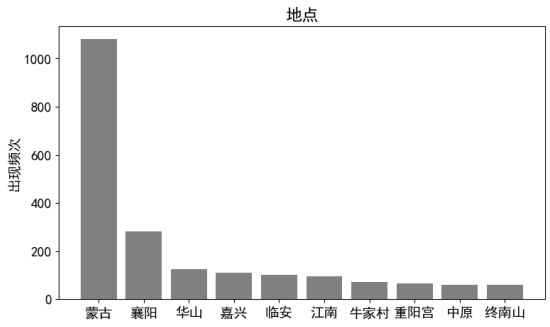

图 2. 高频地点实体

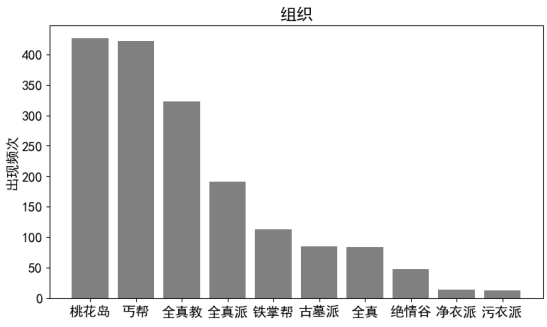

图 3. 高频组织实体

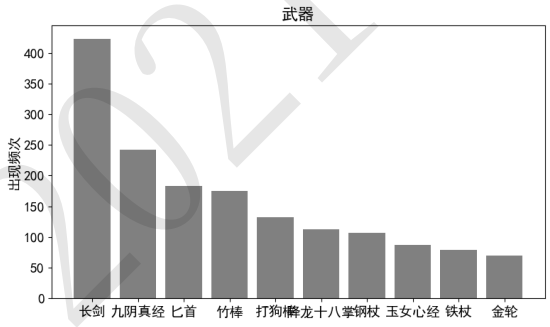

图 4. 高频武器实体

3.6 标注一致性分析

为保证标注质量，本次标注工作采用多轮迭代修正的模式进行标注规范的修订和标注工作，每一篇小说文本同时有两名标注人员进行标注。首先由一标负责人对小说文本进行初步标注，得到一标结果；然后由二标负责人对一标的标注结果进行检查验证，得到二标结果。如有一标二标标注不一致的地方，二人进行讨论并给出解决方案，最后再由一标负责人对标注结果进行确认，得到三标结果。对于标注过程分歧较多的情况，进行了反复讨论。

本次标注使用F1值 (George and Rothschild, 2005)作为一致性评价标准，分别对四种实体进行标注一致性检验。由表6可以看出，不同实体的标注一致性存在差异，其中人物和组织实体标注一致性较高，武器和地点实体一致性较低。总体标注一致性微平均达到0.9381，宏平均达到0.8933，表明该数据集是可信赖的 (Ron and Poesio, 2008)。

PER	LOC	ORG	WEP	min-F1	mac-F1
0.9515	0.8627	0.9032	0.8559	0.9381	0.8933

表 6. 标注一致性分析

4 融入篇章信息的命名实体识别模型

本文通过融合两个模型来进行文学作品的命名实体识别，模型整体结构如图5所示。首先使用BiGRU-CRF (Cho et al., 2014)模型来预测数据的初步标签，同时将隐藏状态和初步标签存入篇章信息字典，然后将篇章信息字典融入到Transformer模型 (Vaswani et al., 2017)中得到新的预测标签，最后通过比较新的预测标签和初步标签的可信度，得到最终模型预测的结果。模型充分考虑篇章不同行数据之间的远距离依赖关系，两个数据集上的多个实验表明了该模型的有效性。

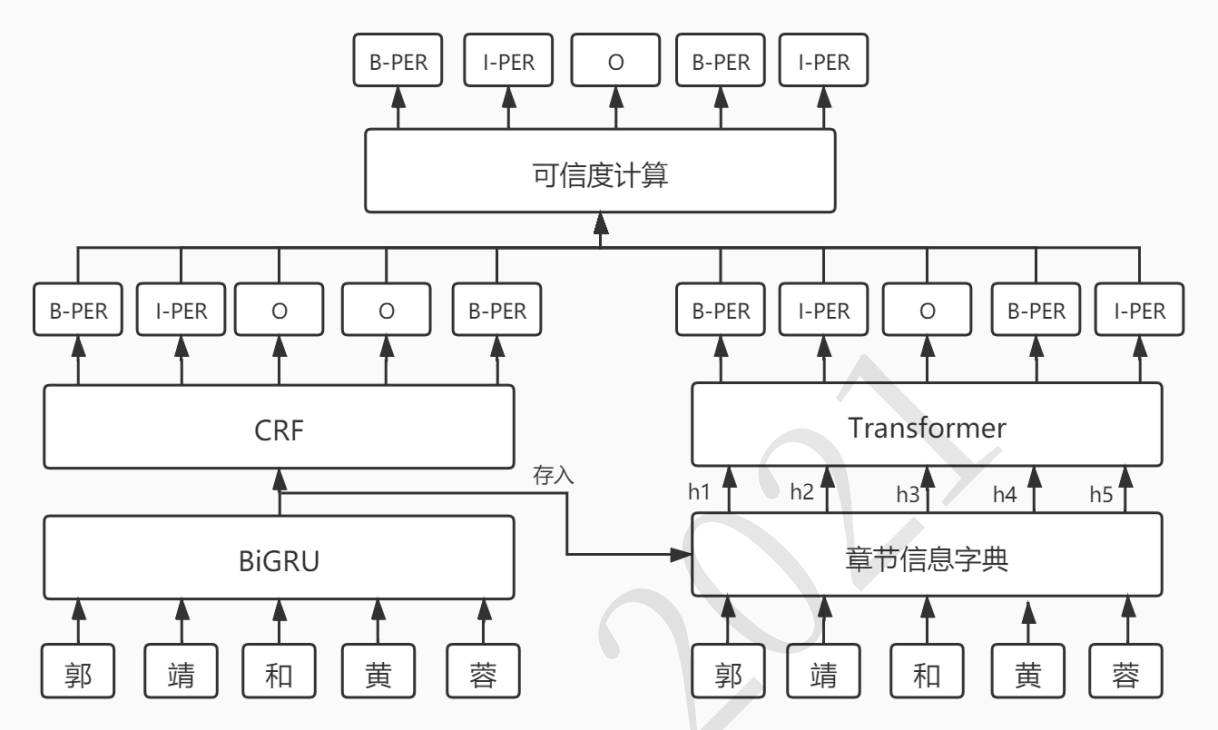

图 5. 模型整体结构图

4.1 篇章信息字典

篇章信息：篇章信息是通过人为对数据集进行划分来构建，篇章篇幅n表示n行连续的数据作为一个篇章。篇章信息字典保存该篇章中所有字符和字符对应的所有的隐藏向量信息，字符i可能在该篇章不同上下文对应多个隐藏向量$\mathrm{h}_{i:1},...\mathrm{h}_{i:n}$，模型在进行预测标签时会同时考虑字符i的所有隐藏向量，从而可以捕捉篇章中不同行数据之间的远程依赖和同行数据之间的近程依赖。

篇章信息字典的构建：模型通过BiGRU+CRF来进行初步预测和篇章信息字典的构建，篇章信息字典表示为$\mathrm{D}=\{\mathrm{ENTRY}_{c1},...,\mathrm{ENTRY}_{cn}\}$,$\mathrm{ENTRY}_{ci}$中包含字符i的隐藏状态向量$\mathrm{h}_i$和对应标签的嵌入$\mathrm{l}_i$，整个字典中包含篇章中所有字符的隐藏状态向量和标签嵌入。由于同一字符i上下文可能不同，因此字符i可以在篇章字典中出现多次（$\mathrm{h}_{i:1}$,$\mathrm{l}_{i:1}$）,...,（$\mathrm{h}_{i:n}$,$\mathrm{l}_{i:n}$），即同一字符i的隐藏状态向量h_i和对应标签的嵌入l_i可能不同。但是对于给定的隐藏状态向量$\mathrm{h}_{i:j}$，只有唯一的标签嵌入$\mathrm{l}_{i:j}$与其对应。

篇章信息字典的应用：在第二阶段Transformer模型中，首先对输入的字符i进行查找篇章字典，得到隐藏状态向量($\mathrm{h}_{i:1},...\mathrm{h}_{i:n}$)，然后通过softmax计算出所有隐藏状态向量的概率分布$\mathrm{p}_{h_{i:j}}$:

$$p_{h_{i:j}} = softmax\left(\mathbf{x}_i^\top W \mathbf{h}_{i:j}\right) \tag{1}$$

其中x_i为字符i的字向量，W为线性运算矩阵，$\mathrm{h}_{i:n}$为对应的隐藏状态向量。然后通过计算每个隐藏状态向量$\mathrm{h}_{i:j}$与该向量的概率$\mathrm{p}_{h_{i:j}}$的乘积得到该隐藏状态向量对字符i的贡献，最后将所有贡

献相加得到字符i的整体隐藏状态向量h_i，h_i包含了整个篇章中字符i的信息。

$$h_i = \sum_{i=1}^{n} p_{h_{i:j}} h_{i:j} \tag{2}$$

4.2 BiGRU-CRF模型

GRU全称为Gated Recurrent Unit，即门限循环单元，是一种改进的循环神经网络RNN (Zaremba et al., 2014)。能够有效解决RNN网络梯度消失和爆炸的问题，同时相较于LSTM模型更为简化。其中z_t为更新门，r_t为重置门，x_t为输入文本的向量，h_t是第t步隐藏状态。

$$z_t = \sigma\left(W^z x_t + U^z h_{t-1}\right) \tag{3}$$
$$r_t = \sigma\left(W^r x_t + U^r h_{t-1}\right) \tag{4}$$
$$\tilde{h}_t = \tanh\left(W^h x_t + U^h\left(h_{t-1} \odot r_t\right)\right) \tag{5}$$
$$h_t = \left(1 - z_t\right) \odot \tilde{h}_t + z_t \odot h_{t-1} \tag{6}$$

CRF全称为Conditional Random Field，即条件随机场，是一种无向图模型 (Lafferty et al., 2001)。该模型能够将随机变量的输入转换为条件概率分布，能够有效学习到训练数据中与标签有关的约束条件，能够保证得到结果的有效性。对于给定数据W=(w_1,...,w_n),其预测标签序列y=(y_1,...,y_n)。计算公式为：

$$P(W, y) = \sum_{i=1}^{n} M_{y_i, y_{i+1}} + \sum_{i=1}^{n} N_{i, y_i} \tag{7}$$

M为条件转移矩阵，$M_{yi,yi+1}$为从yi标签转移到yi+1标签的概率M[yi,yi+1]，Ni,yi表示第i个字被标记为标签yi的概率。

4.3 Transformer模型

Transformer模型首先通过多头自注意力机制对融入篇章信息的隐藏状态向量h进行注意力计算，第一步先对Q矩阵和K矩阵进行相似度计算，得到相似度f：

$$f(Q, K_i) = Q^{\top} K_i \tag{8}$$

相似度f通过softmax函数归一化后与相应的V矩阵加权求和得到最后的注意力：

$$Attention(Q, K, V) = \sum_{i} softmax(f(Q, K_i)) V_i \tag{9}$$

Transformer模型通常使用多头注意力，即模型通过将d_model维的Q，K，V矩阵分别进行h次线性变换，将Q,K,V映射到d_model/h维，再分别进行注意力的计算，最后将这些注意力连接得到结果,其中W_i^Q,W_i^K,W_i^V为投影矩阵：

$$MultiHead(Q, K, V) = Concat(head1, head2, ..., headn) W^O \tag{10}$$
$$headi = Attention(QW_i^Q, KW_i^K, VW_i^V) i = 1, 2, 3..., h \tag{11}$$

4.4 可信度计算

模型最终的输出结果通过计算标签预测可信度来确定，标签可信度计算过程如下：首先使用相同的方法分别计算两个模型预测标签的错误率r_i：

$$r_i = -\sum_{j=1}^{n} p_j log p_j \tag{12}$$

其中p_j表示字符i预测为标签j的概率，对于输入序列X=(x1,x2,...,xn)，两个模型的错误率分别为R1=(r_1^1,r_2^1,...,r_n^1)，R2=(r_1^2,r_2^2,...,r_n^2)，然后分别比较对应位置错误率，错误率低的可信度高，作为模型的最终输出结果。

5 实验结果

5.1 实验设置

实验使用哈工大LTP工具 (Che et al., 2020)，BiLSTM-CRF，BiGRU-CRF，Lattice-LSTM-CRF和BERT-LSTM-CRF模型作为基线模型，其中LTP工具不训练直接用于识别。主要进行了以下四个实验，实验一：整体数据集划分训练集：验证集：测试集为2：1：1进行实验。实验二：对训练数据中未出现的新实体（OOV）召回率单独进行测评。实验三：针对不同篇章篇幅设置进行实验，篇章篇幅设置为n表示n行连续数据做为一个篇章，实验得到最佳篇章篇幅设置。实验四：由于98年人民日报1月命名实体识别数据集具有天然的篇章信息（每一篇新闻为一个篇章），对该数据集划分训练集：验证集：测试集3：1：1，分别进行保留原有新闻篇章和去除篇章的实验，进一步验证篇章信息的有效性。实验超参数设置如表7所示。

参数	参数值
learning_rate	0.015
batch_size	1 document
epochs	100
hidden_dim	400
GRU_layers	1
dropout	0.25
transformer_layers	3
n_head	6
optimizer	SGD AdamW
clip_grad	1

表 7. 超参数设置

5.2 实验结果及分析

实验一结果如表8所示，其中BiLSTM-CRF，BiGRU-CRF，Lattice-LSTM-CRF和本文提出的模型使用同一个字向量表，可以看到本文提出的模型明显优于这三种模型。LTP工具由于其训练数据为通用领域，在文学领域数据集上结果较差，Lattice-LSTM-CRF效果低于BiLSTM-CRF的原因可能是因为Lattice-LSTM模型中的词向量表与本数据集领域差异较大。BERT-LSTM-CRF模型由于引入外部预训练数据，宏平均mac-F1高于本文提出模型，但其微平均mic-F1值低于本文提出模型。由此可知本文所提出的模型在不引入外部资源的情况下，效果达到了最好。

model	PER	ORG	LOC	WEP	mic-F1	mac-F1
LTP	65.81	56.02	0	0	61.21	30.46
BiGRU-CRF	93.55	86.72	66.38	41.46	88.94	72.03
BiLSTM-CRF	92.70	81.86	57.97	52.54	88.50	71.27
Lattice-LSTM-CRF	84.57	77.76	66.67	50.98	81.56	70.00
BERT-LSTM-CRF	**97.87**	95.47	**84.37**	80.51	95.74	**89.56**
Our-model	97.81	**95.58**	82.41	**82.28**	**95.89**	89.52

表 8. 整体实验结果

实验二结果如表9所示，模型对于OOV实体的识别效果均低于实验一。Lattice-LSTM在实验二中的识别OOV的能力优于BiLSTM-CRF和BiGRU-CRF。BERT-LSTM-CRF模型mac-Recall高于本文提出模型，但由于其人物实体识别率较低且人物实体占比较大，故mic-Recall远低于本文提出模型。

model	PER	ORG	LOC	WEP	mic-Recall	mac-Recall
BiGRU-CRF	30.6	9.63	5.20	11.02	17.40	14.11
BiLSTM-CRF	32.84	16.67	6.00	11.81	19.87	16.83
Lattice-LSTM-CRF	33.63	5.32	30.95	29.25	30.95	24.79
BERT-LSTM-CRF	38.06	**50.00**	**39.22**	36.22	37.74	**40.88**
Our-model	**57.33**	10.00	22.08	**53.24**	**49.60**	35.66

表 9. OOV实体识别效果

实验三结果如图6和图7所示，document为n表示一个篇章中包含n个句子。通过图片可以看到，随着篇章篇幅的增大，模型识别效果表现出先提升再下降的趋势，并在篇章篇幅为20的时候达到最好。原因可能是在篇章篇幅过小时，篇章信息过少，模型只能学到有限的篇章信息；而在篇章篇幅过大时，篇章内数据之间关联性下降，模型可能学到有偏误的篇章信息。以上分析证明了篇章信息在模型进行命名实体识别中是有效的。

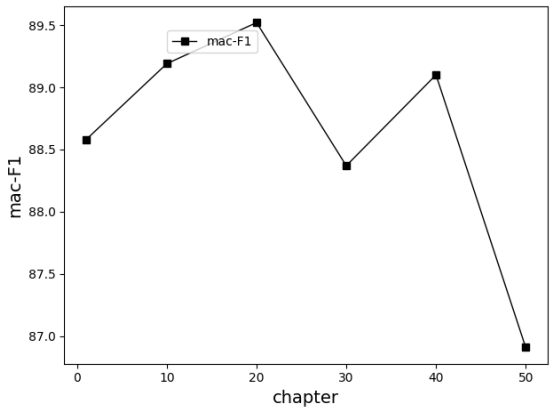

图 6. 不同篇章篇幅mac-F1值

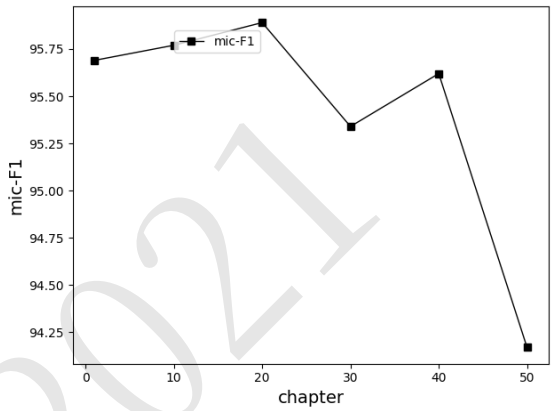

图 7. 不同篇章篇幅mic-F1值

实验四结果如表10所示，人民日报数据集具有天然的篇章信息，同一篇新闻报道往往主题相同，实体之间联系较为紧密，模型能够学习到的篇章信息较多。实验结果也证明了这一点，保存篇章信息比去除篇章信息的平均F1值提高了3%以上。

model	PER	ORG	LOC	mic-F1	mac-F1
LTP	**94.73**	42.00	72.87	75.64	69.87
Our-model w/o doc	82.38	87.94	83.76	83.66	84.69
Our-model	86.89	**93.00**	**85.84**	**86.99**	**88.58**

表 10. 人民日报数据集实验结果

5.3 案例分析

案例分析如图8所示，第一句话中“靖哥哥”的“靖”和后面同一篇章另一句话中“郭靖”的“靖”形成篇章级依赖关系，二者均被预测为人物标签。而后面一句话中“郭靖”的“靖”和“靖康之耻”中的“靖”属于本地句子级依赖关系，但二者的标签并不相同。由此可知，如果只考虑句子级的依赖关系，有可能导致模型预测错误，而加入篇章信息之后，可以有效避免这一问题发生。

图 8. 案例分析

错误分析如图9，第一句话和第二句话中出现的“孙婆婆”实体均被错误识别为“孙婆”。原因可能为“孙婆婆”实体出现较少且篇章内出现大量“老太婆”、“老婆婆”等不是命名实体的字符串，“孙婆婆”中的第二个“婆”错误依赖了“老太婆”、“老婆婆”中的“婆”，从而导致“孙婆婆”中的第二个“婆”被标注为标签O。

图 9. 错误分析

6 文学作品命名实体识别的应用

从小说中识别出命名实体之后，就可以进行后续的实体关系的研究。以人物之间的关系为例，根据人物实体在篇章中的共现情况可以构造社会网络，进而可以利用复杂网络的技术来分析人物之间的关联关系。使用模型在《射雕英雄传》上的NER识别结果，篇章大小设置为20句，即20句以内共同出现的人物具有共现关系，构造的社会网络（部分）如图10所示，也可以获取以个人为中心的网络，如图11所示为主要人物“杨康”的人物关系网络。为了获取更精确的社会网络，后续可以在NER基础上进一步做指代消解，以获得实体层面的共现关系。另一方面，也可以考虑用对话关系代替共现关系，探索更加多样性的社会网络 (Jia et al., 2020)。

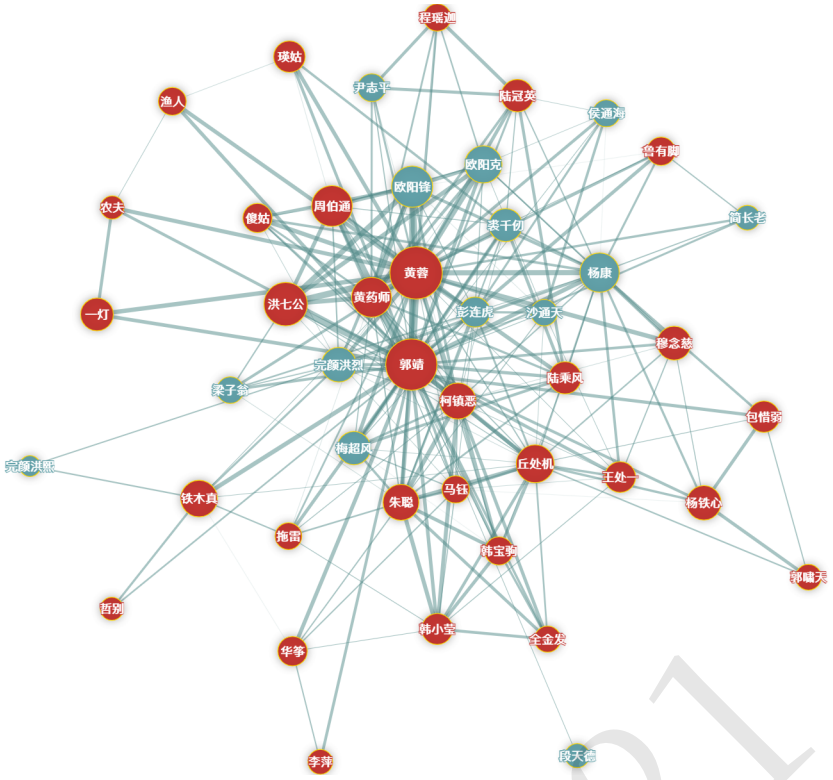

图 10. 《射雕英雄传》人物社会网络

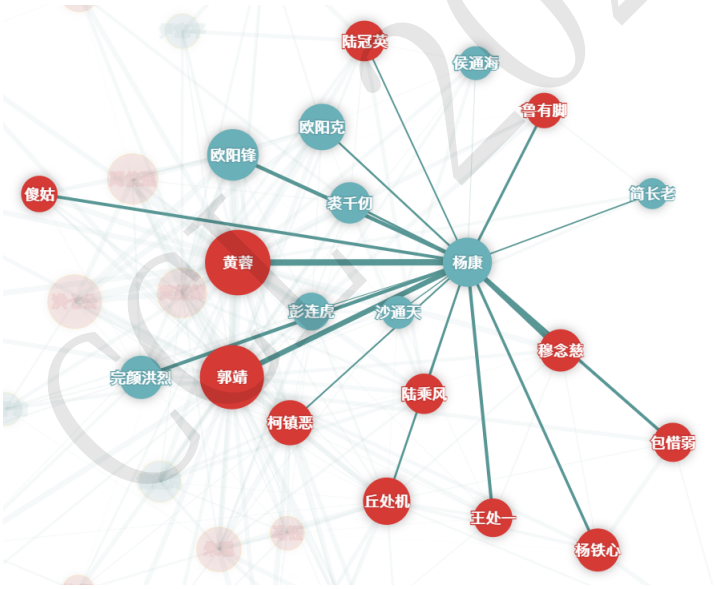

图 11. 《射雕英雄传》杨康社会网络

7 总结与展望

本文基于两部金庸武侠小说构建了命名实体识别数据集，同时提出一种融入篇章信息的命名实体识别模型，通过多角度实验证明了该模型的有效性。最后本文构建了《射雕英雄传》的人物社会网络，体现了文学领域NER的应用价值。

本文构造的数据集还存在规模较小、来源单一的问题，下一步工作将继续加大文学作品数据标注规模，丰富文本来源；优化命名实体识别模型，并探索人物实体之间的关系以及不同实体之间的关系。

第二十届中国计算语言学大会论文集，第600页-第611页，呼和浩特，中国，2021年8月13日至15日。

参考文献

李保利, 陈玉忠, 俞士汶. 2003. 信息抽取研究综述. 计算机工程与应用,039(010):1-5,66.

Vincent Labatut and Bost Xavier. 2019. *Extraction and analysis of fictional character networks: A survey*, volume 52. ACM Computing Surveys (CSUR),52(5): 1-40.

Sims Matthew, Jong H. Park, and David Bamman. 2019. *Literary event detection.* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: 3623-3634.

David Bamman, Brendan O. Connor, and Noah A. Smith. 2013. *Learning latent personas of film characters.* Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics(Volume 1: Long Papers): 352-361.

Matthew L. Jockers. 2013. *Macroanalysis: Digital methods and literary history.* University of Illinois Press.

林峰, 赵广平, 林娜, 吴亚楠. 2018. 《红楼梦》文本的社会网络结构分析. 石家庄铁道大学学报(社会科学版)(2018 年01): 58-63.

韩忠明, 陈炎, 刘雯. 2017. 社会网络节点影响力分析研究. 软件学报,28(1): 84-104.

Anu Thomas and S. Sangeetha. 2020. *Deep Learning Architectures for Named Entity Recognition: A Survey.* Advanced Computing and Intelligent Engineering:215-225.

Ronan Collobert, Jason Weston, Leon Bottou, et al. 2011. *Natural language processing (almost) from scratch.* Journal of machine learning research 12(ARTICLE):2493-2537.

Onur Kuru, Ozan Arkan Can and Deniz Yuret. 2016. *Charner: Character-level named entity recognition*, . Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers:911-921.

柏兵,侯霞,石松. 2018. 基于*CRF* 和*BI-LSTM* 的命名实体识别方法. 北京信息科技大学学报, 33(06):27-33.

Zhang Yue and Yang Jie. 2018. *Chinese NER Using Lattice LSTM.* Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers): 1554-1564 .

Liu Liyuan, Shang Jingbo, Ren Xiang, et al. 2018. *Empower Sequence Labeling with Task-aware Neural Language Model.* Proceedings of the AAAI Conference on Artificial Intelligence: 5253-5260.

王月, 王孟轩, 张胜. 2019. 基于*BERT* 的警情文本命名实体识别. 计算机应用, 40(2): 535-540.

陈茹, 卢先领. 2020. 融合空洞卷积神经网络与层次注意力机制的中文命名实体识别. 中文信息学报, 34(8): 70-77.

Julian Brooke, Timothy Baldwin, and Adam Hammond. 2016. *Bootstrapped text-level named entity recognition for literature*, . Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers):344-350.

Xu Jingjing, Wen Ji, Sun Xu, and Su Qi. 2017. *A discourse-level named entity recognition and relation extraction dataset for chinese literature text.* arXiv preprint arXiv:1711.07010.

Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. 2015. *Mr. bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts.* Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing:769-774.

谢韬. 2018. 基于古文学的命名实体识别的研究与实现. 北京邮电大学.

David Bamman, Sejal Popat, and Sheng Shen. 2019. *An annotated dataset of literary entities.* 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): 2138-2144.

Kyunghyun Cho, et al. 2014. *Learning phrase representations using RNN encoder-decoder for statistical machine translation.* arXiv preprint arXiv:1406.1078.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. 2017. *Attention is all you need.* arXiv preprint arXiv:1706.03762.

John Lafferty, Andrew McCallum, and Fernando C. Pereira. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* Proc. 18th International Conf. on Machine Learning.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. *Recurrent neural network regularization.* arXiv preprint arXiv:1409.2329.

Che Wanxiang, Feng Yunlong, Qin Libo and Liu Ting. 2020. *N-LTP: A Open-source Neural Chinese Language Technology Platform with Pretrained Models.* arXiv preprint arXiv:2009.11616.

Gui Tao, Ye Jiacheng, Zhang Qi, et al. 2020. *Leveraging document-level label consistency for named entity recognition.* aProceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020: 3976-3982.

赵京胜, 张丽, 朱巧明, 周国栋. 2020. 中文文学作品中的社会网络抽取与分析. 中文信息学报, 31(2): 99-106.

Jia Yuxiang, Dou Huayi, Cao Shuai, et al. 2020. *Speaker Identification and Its Application to Social Network Construction for Chinese Novels.* International Journal of Asian Language Processing,30(04), 2050018: 1-18.

Hripcsak George and Adam S. Rothschild. 2005. *Agreement, the f-measure, and reliability in information retrieval.* Journal of the American medical informatics association,12(3): 296-298.

Artstein Ron and Massimo Poesio. 2008. *Inter-coder agreement for computational linguistics.* Computational Linguistics,34(4): 555-596.