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

罗 凌 杨志豪 宋雅文 李 楠 林鸿飞

(大连理工大学计算机科学与技术学院 辽宁 大连 116024)

摘 要 近年来,电子病历文本数据不断增长,这为医学研究提供了丰富的知识来源.结合领域需求,采用有效的文本挖掘技术从电子病历文本中自动快速、准确地获取医疗知识,将对医疗健康领域的研究产生极大的推动作用.中文临床电子病历命名实体识别作为中文医学信息抽取的基本任务,已经受到了广泛关注.目前大多数中文电子病历实体识别工作都是在传统通用的文本表示向量基础上,通过特征工程来提升模型在医疗领域上的性能,缺乏适合中文生物医学特定领域的预训练表示向量.此外,目前现存的中文电子病历标注数据十分稀缺,标注电子病历实体需要具备专业的医学背景知识,且耗时耗力.针对这些问题,本文提出了一种基于笔画ELMo和多任务学习的中文电子病历实体识别方法.首先以笔画序列为输入对ELMo表示学习方法进行改进,利用海量无标注的中文生物医学文本学习上下文相关且包含汉字内部结构信息的笔画ELMo向量,然后构建基于多任务学习的神经网络模型来充分利用现存数据提升模型性能.此外,本文还系统地比较了实体识别常用额外特征(包括词向量、词典和部首特征)以及主流神经网络模型(包括CNN、BiLSTM、CNN-CRF和BiLSTM-CRF模型)在中文电子病历实体识别任务上的性能.实验结果表明,在该任务上BiLSTM-CRF模型获得了比其它模型更好的结果,常用额外特征中词典特征最为有效.相比其它现存方法,本文提出的基于笔画ELMo和多任务学习的神经网络模型在CCKS17和CCKS18 CNER数据集上都获得了更好的结果,F值分别为91.75%和90.05%.

关键词 笔画ELMo;多任务学习;神经网络;实体识别;中文电子病历

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2020.01943

Chinese Clinical Named Entity Recognition Based on Stroke ELMo and Multi-Task Learning

LUO Ling YANG Zhi-Hao SONG Ya-Wen LI Nan LIN Hong-Fei

(*School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning* 116024)

Abstract In recent years, the number of electronic medical record text has grown substantially, which provides a rich source of knowledge for medical research. According to the medical domain demand, effective text mining technology can obtain medical related information from the massive electronic medical records efficiently and accurately, which will greatly promote the research in the medical health field. Chinese Clinical Named Entity Recognition (CNER) is a fundamental task for Chinese medical information extraction, which has received much attention. However, most of the existing Chinese CNER works are based on traditional text representation embeddings (i.e., context-independent representation for each word) and depend on effective feature engineering to improve the performance of models in the medical field. There is less related work in Chinese biomedical pretrained text embeddings. In addition, the existing Chinese CNER

收稿日期:2019-10-10;在线发布日期:2020-02-15.本课题得到十三五国家重点研发计划项目(2016YFC0901900)资助.**罗 凌**,博士,主要研究方向为生物医学文本挖掘、深度学习和自然语言处理.E-mail:lingluo0415@gmail.com.**杨志豪**(通信作者),博士,教授,主要研究领域为文本挖掘、机器学习和自然语言处理.E-mail:yangzh@dlut.edu.cn.**宋雅文**,硕士,主要研究方向为生物医学文本挖掘.**李 楠**,博士,主要研究方向为生物医学文本挖掘和知识图谱.**林鸿飞**,博士,教授,主要研究领域为搜索引擎、文本挖掘、情感计算和自然语言理解.

dataset size is small, and medical entity annotation requires medical background knowledge, which is time-consuming and labor-intensive. To address the problems, this paper proposes a Chinese CNER method based on stroke ELMo and multi-task learning. Firstly, a stroke ELMo (Embeddings from Language Models) model is proposed to obtain Chinese pretrained text representation. The ELMo method is improved by taking the stroke sequence as input. It is a context-dependent representation method and can learn rich structure information of the Chinese characters from the large Chinese biomedical text corpus. To learn high quality Chinese biomedical text representations, the massive Chinese medical abstracts were downloaded from the CNKI website. Then these abstracts and the Chinese electronic medical record texts provided by the China Conference on Knowledge Graph and Semantic Computing (CCKS) challenge were used to train the stroke ELMo embeddings. The experimental results show that stroke ELMo embeddings achieve the better performance than the traditional word2vec embeddings. When the concatenation of the word2vec and stroke ELMo embeddings as input is fed into the model, the model obtains the best performance. Secondly, we explored the effect of multi-task learning on the Chinese CNER task. The single task model, fully-shared multi-task learning model and shared-private multi-task learning model are compared on the CCKS17 and CCKS18 data sets. The experimental results show that the shared-private multi-task learning model achieves the best F-score. It can utilize the correlation of the tasks to improve the model performance and make full use of the existing datasets. We also tested the performance of the multi-task learning model on the different sizes training data sets. The shared-private multi-task learning model trained on only 60% of the training data can achieve better performances than the single task model trained on the complete training data on the CCKS17 and CCKS18 CNER datasets. Moreover, the effects of common NER features (i. e., word embedding, dictionary and radical features) and neural network models (i. e., CNN, BiLSTM, CNN-CRF and BiLSTM-CRF models) were investigated for the Chinese CNER task. The experimental results show that the BiLSTM-CRF model outperforms the other models. Among other features, the dictionary feature is most effective. Finally, compared with other existing methods, our neural network model based on stroke ELMo and multi-task learning achieves better performances on the CCKS17 and CCKS18 CNER datasets (the F-scores of 91.75% and 90.05%, respectively).

Keywords stroke ELMo; multi-task learning; neural networks; named entity recognition; Chinese electronic medical records

1 引 言

近年来,随着大量电子病历的产生以及医疗信息服务和医疗决策支持的潜在要求,医疗信息处理已成为一个研究热点.临床电子病历命名实体识别(Clinical Named Entity Recognition,CNER)作为医疗信息抽取的基础任务,已经受到了广泛关注,并且在国际上已经多次举办了相关评测[1-3].为了推动CNER系统在中文临床文本上的表现,中国知识图谱与语义计算大会(China Conference on Knowledge Graph and Semantic Computing,CCKS)在2017年和2018年都组织了面向中文电子病历的命名实体识别评测任务.该评测任务目标是从给定的电子病历纯文本文档中识别并抽取出与医学临床相关的实体提及,并将它们归类到预定义的类别.图1展示了CCKS18 CNER评测数据的一个样例.

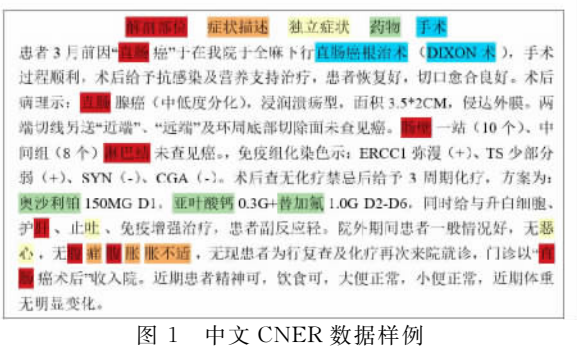

图1 中文CNER数据样例

在中文 CNER 任务上，已经有不少研究方法被提出，这些研究主要集中在对领域特征的探索上，即在通用领域 NER 方法的基础上，加入中文汉字特征和电子病历知识特征等，通过特征工程来提升模型性能[4-7]. 这些方法取得了不错的成绩，但需要人工专家设计大量特征，耗时耗力，且最终实体识别性能受特征影响较大，泛化能力较弱，在不同的语料上需要重新寻找最优的特征组合. 最近，一些基于大规模无标注数据的预训练方法被提出（例如 ELMo[8]、BERT[9]和 GPT[10]），其中 Peng 等人[8]提出的 ELMo (Embeddings from Language Model)模型就是代表工作之一，该工作通过语言模型预训练学习得到上下文依赖的 ELMo 向量，在 6 项 NLP 任务上实验表明，加入 ELMo 后结果都获得了显著提升，具有较强的泛化能力，并无需大量特征工程. 但原始的 ELMo 模型是对英文进行建模，而中文与英文特点不同，如何充分利用大规模无标注数据学习更适合中文生物医学领域的预训练向量表示，进而提升 CNER 的性能是本文的研究重点之一.

此外，中文 CNER 任务还面临的一个主要挑战是带标注的训练数据集稀缺且规模小. 而且不同的机构标注的数据集，其实体类型和标注规则也不相同，无法通过简单地合并多个数据集来扩大数据规模. 以 CCKS 2017 年和 2018 年两个 CNER 数据集（以下简称 CCKS17 和 CCKS18）为例. CCKS17 数据集由北京极目云健康科技有限公司提供，训练集标注了 300 篇电子病历文档，具体实体类型为症状和体征、检查和检验、疾病和诊断、治疗、身体部分，共五类. 而 CCKS18 数据集由北京医渡云技术有限公司提供，训练集标注了 400 篇电子病历文档，具体实体类型为解剖部位、症状描述、独立症状、药物和手术共五类. 两个数据集中的实体类型有一定的交叉，但又不完全相同. 如何利用现存的多个相关小数据集来提升中文 CNER 模型的性能也是本文的一个研究重点.

针对上述这些问题，本文提出了一种基于笔画 ELMo 和多任务学习的中文电子病历命名实体识别方法. 该方法首先在大规模无标注数据上使用双向语言模型预训练得到笔画 ELMo 向量作为输入特征，然后在 CCKS17 和 CCKS18 两个中文电子病历实体识别任务上构建了基于多任务学习的神经网络模型来识别电子病历实体. 本文的主要贡献总结如下：

（1）将 ELMo 模型应用到了中文上，并提出了一种基于汉字笔画的中文 ELMo 模型. 以汉字笔画序列作为输入，通过双向语言模型预训练方式来学习上下文相关且包含汉字内部结构信息的中文 ELMo 向量. 并利用中国知网（http://www.cnki.net/）下载的医学文摘以及 CCKS 发布的电子病历文本，训练了面向生物医学领域的中文 ELMo 向量. 还与传统词向量以及新闻领域的 ELMo 向量进行了对比. 实验结果表明，本文提出的生物医学领域笔画 ELMo 向量在 CNER 任务上比其它向量获得了更好的结果.

（2）探索了多任务学习在中文 CNER 任务上的效果，在 CCKS17 和 CCKS18 两个数据集上对比了单任务模型、全共享多任务模型和私有共享多任务模型. 并在不同大小的训练数据集上测试了多任务学习模型的性能. 实验结果表明私有共享多任务学习框架效果最好，能够充分利用两个任务的相关性，进一步提升模型性能. 在小数据集上，多任务模型也获得了不错的表现.

（3）系统地对比了目前主流实体识别模型（包括 CNN、BiLSTM、CNN-CRF 和 BiLSTM-CRF 模型）在中文 CNER 任务上的效果，还调查了一些现存中文向量表示方法和常用特征（词典、中文部首和词向量特征）对模型的性能影响.

最后实验表明，在 CCKS17 和 CCKS18 CNER 语料上，本文方法都取得了较好的结果（最佳模型 F 值分别为 91.75%和 90.05%）.

2　相关工作

命名实体识别（Named Entity Recognition，NER）作为信息抽取的一项基本任务，已经有很多方法被提出. 在英文的 NER 任务上，基于神经网络的方法已经成为了目前主流方法，其中将神经网络和条件随机场结合的 CNN-CRF 模型[11-12]和 BiLSTM-CRF 模型[13-15]最具代表性. 与英文表达不同，中文文本中并没有明确的边界信息. 所以根据文本的切分方式，中文 NER 方法大体可以分为两类：基于词的方法和基于字的方法. 对于基于词的方法，首先利用中文分词工具进行分词，然后再进行实体识别[16-17]. 所以词边界也是实体的边界. 这类方法主要存在错误传播的问题，因为在中文分词阶段，如果由于分词错误将一个实体的边界进行了错误的切分，那么后续的 NER 将无法正确识别这个实体. 对于基于字的方法，不进行分词，而是将每个字进行切

分,然后对字序列进行标注.这类方法不存在分词错误传播的问题,但其主要缺点是无法充分精确地利用词信息.所以在基于字的模型上,研究者们主要关注如何更好地利用词信息[18-20].目前大多数研究表明,由于基于词的方法存在分词错误传播问题,在中文 NER 上,基于字的方法通常优于基于词的方法[21-22].

作为中文 NER 中特定领域的 CNER,目前大多数中文 CNER 研究主要集中在对领域特征的探索上.即在基于字的通用 NER 方法基础上,通过加入汉字和医学词典等特征,来提升模型性能[4-7].例如,Ji 等人[7]直接将 BiLSTM-CRF 模型应用到中文 CNER 上,然后通过药物词典和后处理规则来提升模型性能.Qiu 等人[5]则使用扩张卷积神经网络进行编码,CRF 进行解码来进行中文电子病历实体识别.Wang 等人[6]提出了五种不同的加入词典特征的方法来提升 BiLSTM-CRF 模型在中文 CNER 上的性能.Yang 和 Huang[4]开发了一个基于丰富特征的 CRF 模型,包括字向量、词性、部首、拼音、词典和规则等丰富特征,在 CCKS 2018 CNER 评测上获得了第一名.这些方法需要进行特征工程,寻找丰富有效的特征来提升模型性能.

随着深度学习的研究热潮,在 NLP 领域中,词向量研究受到了广泛关注,且取得了巨大成功.目前用于学习词向量的大多数模型基于分布式假设基本原则[23]:类似的词语往往出现在类似的语境中.其中,word2vec[24]和 GloVe[25]是词向量的两个代表性模型,在 NLP 任务中被广泛使用.由于中文的特殊性,一些工作也专门研究了中文词的表示方法.Yang 和 Sun[26]使用了一个中文同义词词典来缓解词或字的多义问题.Xu 等人[27]利用翻译工具从其它语言中提取语义知识,以捕获单词中字的语义信息.这些方法主要通过外部知识来提高词向量质量,此外研究者们也提出了一些利用中文内部结构信息的方法.Cheng 等人[28]提出了 CWE 模型,用字和词的联合学习充分利用字级别特征信息,提高中文词向量的质量.Xin 和 Song[29]提出了一种联合嵌入中文词、字和更细粒度字子成分的方法 JWE.Cao 等人[30]提出了 cw2vec 中文词向量,他们利用词的 n 元笔画信息来学习每个字之间的内部结构联系.但是这些工作基本都是在传统词向量的基础上针对中文特点进行改进,这些模型仍是基于词的模型,且都是上下文无关的词向量.最近,一些基于大规模无标注数据的预训练方法被提出(例如 ELMo[8]、BERT[9]和 GPT[10]),其中 Peters 等人[8]提出的 ELMo 模型就是比较有代表性的工作.ELMo 使用海量无标注文本对语言模型进行预训练,然后通加权求和每层神经网络的中间表示作为上下文相关的词向量表示.它根据每个单词的上下文生成该单词的词向量,从而允许对同一个词的不同的语义进行表示.

最近,多任务学习(Multi-Task Learning,MTL)已经开始成功地应用于 NLP 领域,旨在利用多个相关任务来促进原始任务的提升.例如在通用新闻领域,Liu 等人[31]利用了 16 个不同的文本分类任务来进行多任务学习,取得了比单任务更好的效果.在生物医学 NER 领域,MTL 也已经开始被探索.Zhao 等人[32]提出了一种多任务学习框架将医学实体识别和标准化进行联合学习.Crichton 等人[33]利用了 15 个生物医学 NER 数据集进行多任务学习.实验表明多个 NER 相关任务通过多任务学习框架可以相互促进提升.为了缓解中文 CNER 标注数据稀缺的问题,本文对中文 CNER 进行了多任务学习的探索,通过利用 CCKS17 和 CCKS18 两个 CNER 任务来学习任务间的相关性,进一步提升模型性能.

3 方 法

本节首先描述我们提出的笔画 ELMo 模型,然后介绍实验测试用的额外特征,最后阐述基于多任务学习的实体识别神经网络框架.

3.1 基于笔画的中文 ELMo 模型

近年来,分布式词向量表示受到了研究者们的广泛关注,特别是在基于深度学习的方法中,词向量通常被用于神经网络模型的输入.而词向量的质量通常通过其编码语法信息和语义信息的能力来衡量,对模型性能有重要影响.

传统的词向量方法(如 word2vec[24]和 GloVe[25])对于词汇表中的每个词仅使用一个全局的向量进行表示,即不同上下文的同一个词都是同一个向量表示,是上下文无关的.但是在实际的文本表达中,不同上下文的同一个词可能会有完全不同的语义.例如考虑下面两个句子:(1)“术后行多西他赛+卡铂方案化疗 3 周期.”(2)“曾在外院及我院行多次化疗.”根据上下文语境可以知道,句子(1)中的“多”是药物“多西他赛”名字的一部分,而句子(2)中的“多”是指不止一次的意思.但是传统的词向量方法训练得到的中文字向量,对于“多”只会使用一个字向量

表示，无法对同一个字的不同语义进行表示. 为了克服传统词向量的这个问题，Peters 等人[8]提出了一种上下文相关的词向量表示方法 ELMo，并在 NLP 多项任务上取得了优异结果.

原始的 ELMo 是应用在英文上，英语文本中词之间天然由空格切分，模型将每个英文单词作为输入单元进行训练. Che 等人[34]将 ELMo 应用到了多种语言上，其中也包括中文. 由于中文文本原本并没有空格对词进行切分，所以他们先使用分词工具对中文文本进行分词，然后利用 ELMo 模型训练出中文 ELMo 词向量. 但由于受分词效果的影响，加入中文 ELMo 词向量对他们的实验任务并没有取得显著的提升. 与 Che 等人不同，考虑到后续 NER 模型是基于字的模型，我们设计了两种 ELMo 模型来学习中文字向量：字符 ELMo(char-ELMo)和笔画 ELMo(stroke-ELMo). 两模型结构分别如图 2(a)和(b)所示.

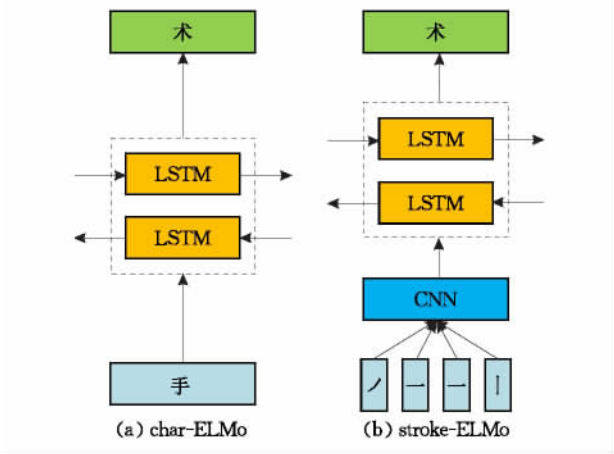

图 2　中文 ELMo 模型结构图

正如相关工作里所述，基于笔画的汉字表示学习，在传统词向量基础上已经有相关研究[30]，他们通过笔画来捕获汉字的内部结构关联，但这些表示方法仍是上下文无关的. 而在最近上下文相关的文本表示方法中，并没有针对汉字特点进行学习，无法捕获汉字内部的结构信息. 因此我们在 ELMo 表示学习基础上提出了笔画 ELMo，通过引入笔画序列来学习到一种即能够捕获汉字的内部结构信息，又是上下文相关的汉字向量表示. 相比字符 ELMo，笔画 ELMo 的主要优势在于：(1) 字符 ELMo 需以字符级别特征为输入，导致低频字很难学习到准确的字向量，而未登录字一般只能通过随机向量来表示. 但是笔画 ELMo 可以从笔画序列生成字向量，缓解了上述低频字和未登录字存在的问题；(2) 语义上相关的汉字常常在内部结构上也有所相关. 例如，“森”、“林”和“木”这三个字，他们分别由不同数量的“木”构成. “呕”、“吐”都与“口”相关，“烧”、“炎”都与“火”相关，等等. 字符 ELMo 无法捕获汉字这样的内部结构信息，而笔画 ELMo 通过构成汉字最基本的笔画序列信息作为输入，在大规模语料上进行语言模型预训练来学习这种内在的关系表示. 下面我们将介绍具体的模型方法.

首先定义第 k 个字的输入表示为 $\boldsymbol{x}_k^{\mathrm{LM}}$，字符 ELMo 和笔画 ELMo 的区别在于使用了不同的输入表示 $\boldsymbol{x}_k^{\mathrm{LM}}$. 对于字符 ELMo，直接将 ELMo 应用在中文字上，即 $\boldsymbol{x}_k^{\mathrm{LM}}$ 为第 k 个字的字向量. 而对于笔画 ELMo，我们利用笔画序列信息作为输入. 具体地，首先利用汉典网站(https://www.zdic.net/)，获取语料库字汇表中每个字的笔画序列信息. 即通过汉典网站构建了一张整个语料库中汉字到其笔画序列的映射表. 对于输入的每个汉字可以通过查找该表获取其笔画序列，例如图 2 中输入的“手”字，根据映射表可以得到其笔画序列为“丿 一一 丨”. 然后将该笔画序列输入到一个卷积神经网络结构中. 在该卷积神经网络结构中，使用了 7 个不同的卷积层，其卷积窗口和卷积特征核大小分别对应为[[1,32],[2,32],[3,64],[4,128],[5,256],[6,512],[7,1024]]，每个卷积层后接一层全局最大池化层得到该卷积特征向量，最后将这些特征向量进行拼接作为这个字的最终向量表示 $\boldsymbol{x}_k^{\mathrm{LM}}$.

在得到字向量之后，将其输入到一个双向语言模型. 这个双向语言模型包含了一个前向语言模型和一个后向语言模型. 在前向语言模型中，定义第 k 个字经过 L 层 LSTM 后得到每层的隐层表示为 $\overrightarrow{\boldsymbol{h}}_{k,j}^{\mathrm{LM}}$(其中 $j=1,2,\cdots,L$). 在 $k-1$ 时刻，该前向语言模型通过前面观测到的字 $t_1,t_2,\cdots,t_{k-1}$ 来预测下一时刻 k 的字 t_k. 即前向语言模型是对一个字序列的联合概率进行建模：$p(t_1,t_2,\cdots,t_N)=\prod_{k=1}^{N}p(t_k\mid t_1,t_2,\cdots,t_{k-1})$. 而反向语言模型与前向语言模型类似，具有相同网络结构，不同在于通过观测未来的字序列对当前时刻进行预测：$p(t_1,t_2,\cdots,t_N)=\prod_{k=1}^{N}p(t_k\mid t_{k+1},t_{k+2},\cdots,t_N)$. 最后合并两个方向的模型，目标函数为最大化两个方向的对数似然函数.

然后在大规模语料上对上述语言模型进行预训练. 对于后续具体任务中每个中文字表示，则通过对该语言模型中间层表示进行线性合并来获取最终的 ELMo 字向量. 对于一个 L 层双向语言模型总共包

含 $2L+1$ 层表示：

$$R_k=\{\boldsymbol{x}_k^{\mathrm{LM}},\overrightarrow{\boldsymbol{h}}_{k,j}^{\mathrm{LM}},\overleftarrow{\boldsymbol{h}}_{k,j}^{\mathrm{LM}}\mid j=1,2,\cdots,L\}$$
$$=\{\boldsymbol{h}_{k,j}^{\mathrm{LM}}\mid j=0,1,\cdots,L\} \quad (1)$$

其中，当 $j=0$ 时，$\boldsymbol{h}_{k,0}^{\mathrm{LM}}$ 是字输入表示向量(即 $\boldsymbol{x}_k^{\mathrm{LM}}$)；此外，$\boldsymbol{h}_{k,j}^{\mathrm{LM}}$ 是两个方向的 LSTM 层表示拼接$[\overrightarrow{\boldsymbol{h}}_{k,j}^{\mathrm{LM}},\overleftarrow{\boldsymbol{h}}_{k,j}^{\mathrm{LM}}]$. 最后的 ELMo 向量通过式(2)得到：

$$\mathbf{ELMo}_k=\gamma\sum_{j=0}^{L}s_j\boldsymbol{h}_{k,j}^{\mathrm{LM}} \quad (2)$$

其中，s_j 是经过 softmax 后标准化的权重，用于组合不同层的表示. γ 是一个参数，有助于优化特定任务的 ELMo 表示.

3.2 额外特征

为了调查其它 NER 常用特征对模型的性能影响，除了上述的 ELMo 向量，我们对以下特征也进行了实验.

(1) 中文字向量. 目前，已经有不少中文分布式表示方法被提出，为了对比现存中文向量表示方法在 CNER 上的效果，我们使用 word2vec[24]，JWE[28] 和 cw2vec[30] 方法分别训练了不同的中文字向量作为对比.

(2) 中文词向量. 由于基于字的模型未充分考虑词信息，我们先使用结巴分词工具对无标注数据进行了分词，然后再使用 cw2vec 方法训练了 50 维的词向量作为特征. 具体使用时，在每个字向量上拼接其词向量.

(3) 词典特征. 在之前的研究中，词典特征已经被验证为有效特征. 我们利用搜狗输入法词典(https://pinyin.sogou.com/dict/detail/index/270)构建了药物词典. 并将训练集中的药物实体也作为词条加入到词典中. 我们将电子病历文本和词典中药物词条进行前向最大长度匹配，然后根据匹配结果使用 BIOES 标签进行编码生成特征. 最后将其随机初始化成一个 50 维的词典特征向量. 我们也尝试了其它类别词典，但由于词典质量问题，除了药物词典，其它类型词典效果并不理想.

(4) 部首特征. 在中文 NER 中，一些中文语言学特征也被用于提升模型性能. 例如部首特征，一些语义相关的字词，常具有相同的部首. 为了验证部首特征在 CNER 上的效果，我们首先从汉典网站获取了每个字的部首信息，然后将其随机初始化成一个 50 维的部首向量作为特征.

3.3 基于多任务学习的神经网络模型

本小节主要介绍我们使用的神经网络模型，首先介绍基础的单任务模型，然后针对 CNER 单数据集规模小的问题，本文利用 CCKS17 和 CCKS18 数据集探索了多任务学习在 CNER 任务上的效果. 具体包括全共享多任务模型和私有共享多任务模型.

3.3.1 单任务模型

对于单任务模型我们主要使用了目前主流的神经网络结合条件随机场模型，模型整体结构如图 3(a)所示. 根据中间的神经网络层不同，分为 BiLSTM-CRF 模型和 CNN-CRF 模型(在实验中，我们也测试了 BiLSTM 和 CNN 模型，即将 BiSLTM-CRF 和 CNN-CRF 模型的 CRF 层替换成了 softmax 层直接进行标签分类，但模型性能有明显下降). 两个模型都采用 BIOES 标签策略对每个中文字预测标签，来进行实体识别.

(1) BiLSTM-CRF. 在本文使用的 BiLSTM-CRF 中，首先，句子通过特征 Embedding 层表示为中文字向量序列，然后输入到 BiLSTM 层. 在 BiLSTM 层中，前向 LSTM 从左到右计算序列的表示，而另一个后向 LSTM 反向计算相同序列的表示. 然后通过连接其左右上下文表示来获得当前字最终的 BiLSTM 层表示. 再使用 Tanh 层来学习更高级别的特征. 最后，使用 CRF 层来预测所有可能标签序列中的最佳标签序列.

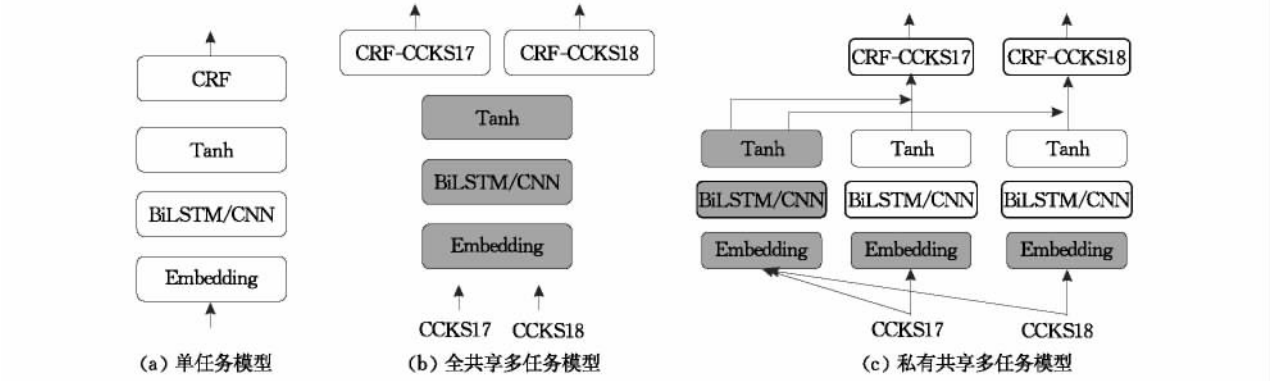

图 3 CNER 神经网络模型结构图(阴影部分表示模型参数共享)

(2) CNN-CRF. 除了 BiLSTM-CRF 模型，CNN-CRF 模型也在 NER 任务中有广泛应用. 与 BiLSTM-CRF 模型主要的不同之处在于，CNN-CRF 模型在神经网络层利用一层卷积层代替了 BiLSTM 层. 通过卷积的滑动窗口来学习当前字的上下文.

3.3.2　全共享多任务模型

在全共享多任务模型(Fully-Shared MTL model，FS)中，除了最后的输出层之外，模型其它所有参数都是共享的，模型结构如图 3(b)所示，图中阴影部分表示模型参数共享. 每个任务都有一个特定任务的 CRF 输出层，该层使用最终的 Tanh 共享层产生的表示进行预测. 该模型通过底层参数全共享来学习两个任务之间的相关性.

3.3.3　私有共享多任务模型

在上述全共享多任务模型中，所有的文本都是通过同一个全共享的神经网络层进行编码，这样的结构特点在于模型参数较少，但是所有任务的信息都只能通过这一个共享神经网络进行表示. 为了进行对比，本文还构建了私有共享多任务模型(Shared-Private MTL model，SP)来进行 CNER 任务. 该模型结构如图 3(c)所示，除了一个共享的神经网络(包括 Embedding 层，BiLSTM/CNN 层和 Tanh 层)之外，每个任务还各自具有一个特定于任务的神经网络(包括 BiLSTM/CNN 和 Tanh 层). 然后将这共享部分和特定于任务的私有部分两者的输出进行拼接作为文本的最终表示，再输入到特定于任务的 CRF 层进行实体识别. 不同于全共享多任务模型，私有共享模型的这种结构能够选择地利用共享信息和特定于任务的信息.

4　实验与分析

4.1　实验设置

我们在 CCKS17 和 CCKS18 两个 CNER 评测数据集上进行实验，来验证本文方法在中文 CNER 任务上的有效性. 这两个数据集都是电子病历文本数据，主要区别在于标注的实体有所不同. 使用多个数据集也是为了验证本文方法具有较好的鲁棒性.

(1) CCKS17. 原始数据集分为训练集和测试集，其中训练集包括 300 个医疗记录，人工标注了五类实体(包括症状和体征、检查和检验、疾病和诊断、治疗、身体部位). 测试集包含 100 个医疗记录；

(2) CCKS18. 同样原始数据集包括训练集和测试集. 其中训练集包括 600 个医疗记录，人工标注了五类实体(包括解剖部位、症状描述、独立症状、药物、手术). 测试集包含 400 个医疗记录. 表 1 列出了数据集中不同类别的实体统计. 由于我们基于多任务的方法涉及到两个任务的交互，需要关注的一个问题是，一个数据集的训练数据与另一数据集中的测试数据之间是否存在明显的重叠现象，因为这会使模型在评价多任务学习时不准确. 经过对比统计，我们发现在医疗记录篇章级别，CCKS17 和 CCKS18 并没有重叠数据；在句子级别，CCKS17 测试集与 CCKS18 训练集重叠率为 0.04%，CCKS18 测试集与 CCKS17 训练集句子重叠率为 0.15%. 为了更准确的评价本文的方法，我们将这些少量的重叠数据从两个数据的训练集中直接去除. 在实验中，我们分别随机选择 20%的训练集数据作为各个数据集的开发集来调整超参数. 此外，为了获得更高质量的预训练字和词向量，我们在知网上下载了医学类文摘，并将 CCKS 提供的中文电子病历文本进行合并，总计 1568458 篇文档作为无标注数据. 为了公平比较，实验中所有字和词向量均使用该数据进行预训练.

表 1　CCKS CNER 语料的数据统计

CCKS17	症状体征	检查检验	疾病诊断	治疗	身体部位
训练集	7831	9546	722	1048	10719
测试集	2311	3143	553	465	3021
CCKS18	解剖部位	症状描述	独立症状	药物	手术
训练集	9472	2484	3712	1221	1329
测试集	6339	918	1327	813	735

在模型参数方面，ELMo 模型使用了默认的网络参数设置，训练迭代 3 次. 对于 CNER 模型，通过在开发集上使用随机搜索方法[35]选择模型超参数. 使用 Adam 算法进行模型参数优化，通过在验证集上的早停策略[36]选择模型训练迭代数. CNER 模型的主要超参数为：LSTM 层大小为 400；CNN 层卷积窗口大小为 3，卷积核大小为 200；Tanh 层大小为 200. 与当时评测相同，本任务在测试集上采用微平均精确率(P)、召回率(R)以及 F 值(F)作为评测指标. 当识别出的实体和人工标注实体边界和类型完全正确才算实体识别正确.

4.2　不同中文字向量性能对比

为了探索目前现存的中文字向量在 CNER 任务上的效果，我们使用 CNN-CRF 模型测试对比了不同维度(包括 50 维、100 维、200 维和 400 维)不同方法的中文字向量在两个 CCKS 数据集上的结果. 比较结果如图 4 所示，其中“rand”表示使用随机初始化字向量；“w2v”表示使用 word2vec 工具中的

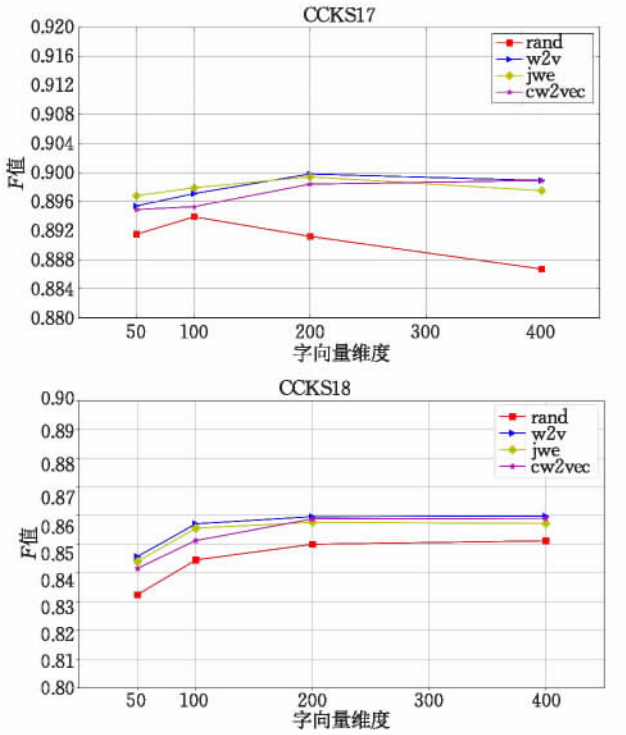

图 4 CCKS 上不同中文字向量性能对比结果

skip-gram 模型[24]训练的字向量；jwe 表示使用 Xin 和 Song[29]提出的一种联合嵌入中文词、字和字子成分的方法训练的字向量；cw2vec 表示使用 Cao 等人[30]提出的基于 n 元笔画信息方法训练的字向量.

从实验结果可以看出，在两个数据集上，所有维度的 w2v、jwe 和 cw2vec 预训练字向量的结果都优于随机初始化的中文字向量结果. 但是专门为中文设计的 jwe 和 cw2vec 方法并没有明显优于传统 w2v 的效果. 可能这些方法是在中文词基础上设计的，而用于训练字向量时，和 w2v 一样，都无法解决歧义问题；且应用于 CNER 任务时，模型训练过程中字向量也会随之微调，所以并没有发挥出其优势. 从不同维度上看，除了 rand 字向量，其它字向量随着维度的上升，模型性能基本都是呈现先上升后平稳的趋势. 而 rand 字向量在 CCKS17 上从 100 维后，更高维度的字向量会带来性能的下降. 可能原因是由于 CCKS17 数据集规模比较小，高纬度的随机字向量增加了模型参数导致模型更难训练. 当在 200 维时，w2v 字向量的效果最好，在 CCKS17 和 CCKS18 上 F 值分别为 89.98%和 85.96%. 后续更高的维度并没有带来明显的性能提升. 最后综合性能和效率的考虑，后续实验中的传统字向量都使用 200 维的 w2v 字向量.

4.3 中文 ELMo 字向量对模型性能的影响

为了验证本文提出的中文 ELMo 字向量的性能，我们将 ELMo 和 w2v 向量以不同组合方式在 CNN-CRF、BiLSTM-CRF、CNN 和 BiLSTM 四个模型上都进行了测试（其中 CNN 和 BiLSTM 模型即将 CNN-CRF 和 BiLSTM-CRF 模型的 CRF 层替换成了 softmax 层直接进行标签分类），CCKS 上结果如表 2 所示. 其中，w2v 表示仅使用 200 维 word2vec 字向量；char-ELMo 表示仅使用 256 维的字符 ELMo 向量；stroke-ELMo 表示仅使用 256 维的笔画 ELMo 向量；w2v + char-ELMo 表示将 word2vec 向量和字符 ELMo 向量进行拼接一起使用；w2v+stroke-ELMo 表示将 word2vec 向量和笔画 ELMo 向量进行拼接；此外，我们还设计了 stroke 和 w2v+stroke 来进行对比，其中，stroke 表示不经过语言模型预训练，直接使用一个 BiLSTM 从笔画序列中学习该字的特征表示向量作为模型输入，w2v+stroke 则表示该笔画特征向量和 word2vec 向量拼接.

表 2 CCKS 上中文 ELMo 字向量性能结果

语料	特征	CNN-CRF			BiLSTM-CRF			CNN			BiLSTM		
		P/%	R/%	F/%	P/%	R/%	F/%	P/%	R/%	F/%	P/%	R/%	F/%
CCKS17	w2v	91.24	88.75	89.98	90.01	90.26	90.13	83.17	84.57	83.87	88.65	88.11	88.38
	stroke	86.37	80.80	83.49	89.03	82.56	85.67	76.09	71.48	73.72	84.87	79.02	81.84
	char-ELMo	90.26	90.06	90.16	89.54	91.40	90.46	87.74	88.28	88.01	89.83	90.50	90.16
	stroke-ELMo	89.75	90.94	90.34	89.81	91.82	90.80	89.19	88.39	88.79	90.59	90.17	90.38
	w2v+stroke	89.77	90.61	90.19	90.68	90.37	90.52	87.88	83.10	85.43	90.30	89.03	89.66
	w2v+char-ELMo	89.44	91.07	90.25	90.43	90.64	90.53	88.70	88.67	88.69	90.36	90.45	90.41
	w2v+stroke-ELMo	90.25	90.87	90.56	91.60	90.18	**90.88**	89.58	89.28	89.43	91.16	90.12	90.63
CCKS18	w2v	86.26	85.66	85.96	87.04	85.84	86.43	78.11	76.28	77.19	84.98	83.73	84.36
	stroke	80.67	78.54	79.59	81.33	79.35	80.33	70.07	65.75	67.84	72.70	66.90	69.57
	char-ELMo	87.99	86.23	87.10	86.94	88.79	87.86	83.33	83.36	83.34	86.53	87.15	86.84
	stroke-ELMo	87.01	88.37	87.69	87.74	88.45	88.09	84.27	83.69	83.98	86.61	87.44	87.02
	w2v+stroke	87.07	85.55	86.30	87.11	86.23	86.67	78.62	77.35	77.98	85.36	84.90	85.13
	w2v+char-ELMo	87.64	87.65	87.64	87.77	88.04	87.90	83.92	84.57	84.25	87.46	87.06	87.26
	w2v+stroke-ELMo	87.98	88.17	88.07	88.16	88.37	**88.27**	84.97	83.92	84.44	87.58	87.73	87.65

实验结果表明，在两个数据集上单独使用每种字向量作为模型输入时，字符 ELMo 和笔画 ELMo 向量在四个模型上都获得了比传统 w2v 字向量更好的表现. 主要原因是 ELMo 方法能够学习上下文依赖的字向量表示，这样可以表示相同字的不同语义. 相比字符 ELMo，笔画 ELMo 获得了更高的 F 值，主要原因是笔画 ELMo 通过大规模语料预训练，从汉字的笔画序列中能够学习到字与字之间更丰富的内部结构关联信息，并且还能够缓解未登录词问题. 对比 stroke 的结果可以看到，不经过预训练，直接用笔画序列作为模型输入也可以进行实体识别，但结果并不理想，在四个模型上 F 值都有明显的下降. 这主要原因可能是由于 CNER 标注训练语料的规模较小，不经过预训练直接使用笔画序列作为模型的输入很难准确学习到汉字语义，导致实体识别的效果不佳.

当 w2v 和 ELMo 向量同时使用时，在所有模型上效果都得到了进一步的提升，其中使用 w2v＋stroke-ELMo 特征的 BiLSTM-CRF 模型在 CCKS17 和 CCKS18 上都获得最高 F 值，分别为 90.88％和 88.27％. 相比传统的 w2v 字向量，加入笔画 ELMo 后所有模型平均 F 值在 CCKS17 和 CCKS18 上分别提升了 2.29％和 3.62％. 说明了本文提出的笔画 ELMo 的有效性. 当笔画以特征的形式加入到模型中时（即 w2v＋stroke），四个模型性能也都获得了轻微提升，这也说明了笔画信息的有效性. 但提升效果不如笔画 ELMo 方法，主要原因是通过大规模数据的语言模型预训练，能够更充分地学习到汉字内部结构信息，而 w2v＋stroke 是以特征形式加入笔画信息，没有预训练过程，且该字向量仍是上下文独立的，无法表示多个语义.

此外，从结果上还可以看出具有循环神经网络结构的 BiLSTM-CRF 和 BiLSTM 模型在 CNER 上的效果要分别优于卷积神经网络结构的 CNN-CRF 和 CNN 模型. 主要原因可能是，相比卷积神经网络结构，循环神经网络结构的特点在于更能捕获长距离依赖信息，CNER 任务上大多实体由多个字构成，这需要长距离的依赖信息来正确识别实体的边界. 当去掉 CRF 层时，BiLSTM 和 CNN 模型的性能都出现了明显的下降，尤其是 CNN. 为了分析原因，我们观察分析了仅使用 w2v 字向量的 CNN 和 BiLSM 在 CCKS17 测试集上的 BIOES 预测结果. 我们发现预测结果中出现了较多的不合理标签序列，例如标签“O”后出现“I-症状体征”、标签“I-身体部位”后出现“I-疾病诊断”、“E-症状体征”后出现“I-治疗”等. 主要原因是在 CNER 任务上由于实体类别的种类较多，BIOES 标签策略总共存在 21 种标签，去掉 CRF 层后，CNN 和 BiLSTM 无法学习到这些标签与标签之间的依赖信息，出现大量不合理标签序列，导致模型性能下降. 相比 BiLSTM，CNN 特点在于学习局部特征，无法捕获长距离依赖上下文，所以 CNN 出现了更明显的性能下降. 但是当加入笔画 ELMo 向量后，CNN 和 BiLSTM 模型在两个数据集上分别平均提升了 6.41％和 2.77％，这说明了笔画 ELMo 能够提供丰富的上下文信息，这给模型带来了性能的提升.

除此之外，我们还在 BiLSTM-CRF 模型上将 ELMo 与 w2v 拼接作为输入，探索了不同维度以及使用不同领域语料训练的 ELMo 对模型的性能影响，结果如表 3 所示. 在训练语料领域方面，可以看到当加入使用领域外语料训练的 ELMo 向量时，模型在 CCKS17 数据集上有轻微的提升，但在 CCKS18 数据上性能反而有所下降，这可能是由于不同领域文本表达不一致，以及用语所有不同导致的. 当加入使用领域内语料训练的 ELMo 时，在两个数据集上模型都能获得更好的表现. 在向量维度方面，相比 256 维的 ELMo 向量，512 维的 ELMo 向量获得了更好的结果. 可能高维度的 ELMo 向量包含了更丰富的语义信息.

表 3　CCKS 上不同维度和领域中文 ELMo 向量性能对比

特征	维度	语料领域	CCKS17-F/％	CCKS18-F/％
w2v	200	生物医学	90.13	86.43
＋char-ELMo	256	新闻	90.20	85.47
＋char-ELMo	256	生物医学	90.53	87.90
＋stroke-ELMo	256	生物医学	90.88	88.27
＋char-ELMo	512	生物医学	90.67	88.19
＋stroke-ELMo	512	生物医学	**91.05**	**88.60**

4.4　额外特征对模型性能的影响

本节实验探索额外特征在 CNER 上的效果. 我们使用 BiLSTM-CRF 模型，根据不同初始输入特征设置了两个基线系统：w2v 基线系统表示仅使用 word2vec 字向量作为输入特征；w2v＋ELMo 基线系统表示使用 word2vec 字向量拼接 512 维的笔画 ELMo 向量作为输入特征. 然后在此基础上加入 3.2 节中的额外特征进行比较，表 4 给出了 CCKS17 和 CCKS18 数据集上不同特征组合的结果.

表 4 CCKS 上额外特征对 BiLSTM-CRF 性能的影响

特征	CCKS17						CCKS18					
	w2v			w2v+ELMo			w2v			w2v+ELMo		
	P/%	*R*/%	*F*/%	*P*/%	*R*/%	*F*/%	*P*/%	*R*/%	*F*/%	*P*/%	*R*/%	*F*/%
基线系统	90.01	90.26	90.13	90.69	91.41	91.05	87.04	85.84	86.43	88.39	88.82	88.60
+词典特征	91.50	89.62	90.55	91.62	91.27	91.44	87.31	87.38	87.34	88.70	89.51	89.10
+部首特征	91.62	89.21	90.39	91.09	91.14	91.11	87.25	86.77	87.01	88.52	88.54	88.53
+词向量	90.72	89.88	90.30	90.55	91.52	91.04	87.05	86.84	86.95	87.64	89.42	88.52
+词典+部首特征	91.34	89.82	90.57	91.40	91.57	91.48	87.75	87.56	87.66	88.83	89.52	89.17
+词典+部首+词向量	90.78	90.61	90.69	91.52	91.49	**91.51**	87.84	88.31	88.08	88.84	89.73	**89.28**

对于不使用 ELMo 向量的 w2v 基线模型，单独加入各项额外特征时，两个数据集上实体识别效果都有所提升，其中词典特征效果最好，*F* 值平均提升了 0.67%，这表明字典提供的先验实体信息有助于模型提高性能. 当加入所有额外特征时，模型获得了最佳性能，相比基线系统在 CCKS17 和 CCKS18 上 *F* 值提高了 0.56%和 1.65%. 对于使用笔画 ELMo 向量的基线模型(w2v+ELMo)，在不使用其他额外特征时性能已经超过了加入全部额外特征的 w2v 模型性能. 单独加入各项额外特征时，只有加入词典特征获得了一定性能的提升，在 CCKS17 和 CCKS18 上 *F* 值分别提升了 0.39%和 0.50%，达到 91.44%和 89.10%，而笔画特征和词特征并没有明显变化. 这也说明笔画 ELMo 通过大规模数据的语言模型预训练，能够自动学习到一定的词信息和汉字内部结构的部首信息，但很难自动学习到词典的先验知识. 当加入所有额外特征后模型在 CCKS17 和 CCKS18 上分别获得最高 *F* 值为 91.51%和 89.28%，相比只加入词典特征，仅提升了 0.07%和 0.18%. 这也表明加入笔画 ELMo 向量后，只需加入词典特征即可获得较好的结果，无需大量特征工程.

4.5 多任务学习模型性能对比实验

本节实验探索多任务学习模型在 CNER 任务上的效果. 首先我们对比了 3.3 节中阐述的单任务模型、全共享多任务模型(FS)和私有共享多任务模型(SP)的性能效果. 我们分别使用了 w2v 字向量作为输入特征和使用 w2v 字向量拼接 512 维的笔画 ELMo 向量作为模型输入进行对比. 实验结果如表 5 所示.

表 5 多任务学习模型性能对比

模型	w2v		w2v+ELMo	
	CCKS17-*F*/%	CCKS18-*F*/%	CCKS17-*F*/%	CCKS18-*F*/%
CNN-CRF	89.98	85.96	90.78	88.24
CNN-CRF-FS	90.36	86.50	91.01	88.51
CNN-CRF-SP	90.47	86.98	91.19	88.72
BiLSTM-CRF	90.13	86.43	91.05	88.60
BiLSTM-CRF-FS	90.71	87.20	91.30	89.29
BiLSTM-CRF-SP	**90.90**	**87.88**	**91.50**	**89.56**

从对比结果可以看到，当仅使用 w2v 字向量作为输入时，在两个数据集上，无论是 CNN-CRF 模型还是 BiLSTM-CRF 模型，基于多任务学习的模型结果都优于单任务模型的结果. 相比全共享模型，私有共享模型能够获得更高的 *F* 值. 主要原因可能是因为私有共享结构使模型能够选择地利用共享和特定于任务的信息，更有效地学习相关信息. 当使用 w2v 字向量和笔画 ELMo 向量同时作为输入时，在两个数据集上得到了相似的结果. 其中 BiLSTM-CRF-SP 模型获得了最好的 *F* 值，在 CCKS17 和 CCKS18 上相比单任务的 BiLSTM-CRF 模型，*F* 值分别提升了 0.45%和 0.96%. 这也说明了本文多任务学习模型的在中文 CNER 任务上的有效性.

为了进一步验证本文提出的笔画 ELMo 和多任务学习模型在小规模数据上的有效性，我们在不同大小的训练数据集上测试了上述基于 BiLSTM-CRF 的单任务和私有共享多任务模型的性能. 具体地，我们按照不同的比例分别从原始标注训练集中随机抽取了部分数据作为新的训练集来测试模型的性能，结果如图 5 所示.

在两个数据集上都呈现了相似的结果. 首先，相比 BiLSTM-CRF 基线系统，在不同规模大小的训练集上，加入本文提出的笔画 ELMo 后，*F* 值都所有提升，CCKS17 和 CCKS18 上 *F* 值分别平均提升了 1.33%和 2.37%. 加入 ELMo 后的私有共享多任务模型能够获得更好的性能，相比基线系统在 CCKS17 和 CCKS18 上 *F* 值分别平均提升了 1.81%和 3.09%. 并且单任务模型与相应的多任务模型的 *F* 值的差距也随着数据集的变小而扩大. 这说明了在小规模 CNER 数据集上，通过本文提出的笔画 ELMo 和多任务学习框架可以获得更好的性能. 其次，从图中结果也可以看到，在两个数据集上仅使用 60%的训练数据规模训练的私有共享多任务模型就获得了比使用完整数据训练的单任务模型更好的性能. 这预示了我们不仅可以使用笔画 ELMo 和多任

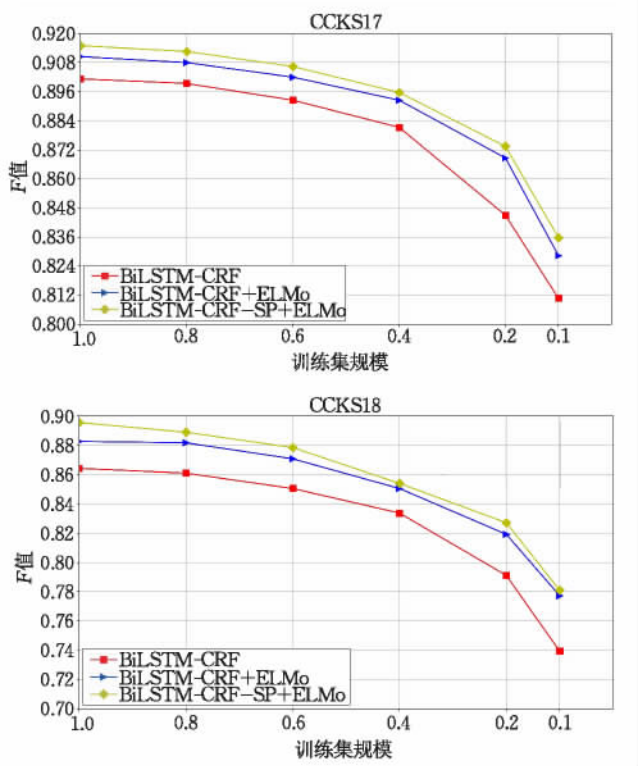

图 5　不同训练集规模对模型性能的影响

务学习提高小数据集上模型的性能，还可以表明在构建新的 CNER 数据集时可以包含更少的注释，减少人工标注的成本. 最后我们还发现，虽然减少训练集的规模会导致模型性能的下降，但是当训练数据集减小到 60%时，即使是单任务的基线系统，模型性能在 CCKS17 和 CCKS18 上 F 值分别也只下降了 0.88%和 1.38%. 直到减少大量数据，在只剩 10%数据量时，模型性能才急剧下降. 这可能原因是，电子病历中的文本描述用语句式比较固定，一些类似的文本表达经常出现，模型能够有效利用相对较少的训练数据来获得足够的性能.

4.6　与其它方法性能对比实验

为了进一步验证本文方法的有效性，我们在 CCKS17 和 CCKS18 两个数据集上分别和其它现存的先进方法进行了比较，并给出了总体以及每一类实体的 F 值结果，如表 6 所示. 对比方法简介如下：

(1) HIT-CNER[37]. 使用字向量作为输入，BiLSTM-CRF 模型识别实体，并使用了无标注数据进行了自训练(Self-training)，该方法当时在 CCKS 2017 评测中获得了第一名.

(2) BiLSTM-CRF-DIC[6]. Wang 等人首先利用相关医学资源构建了医学领域词典，然后以词典特征和字向量作为输入，BiLSTM-CRF 模型识别实体，在 CCKS2017 数据集上取得了不错的结果.

(3) RD-CNN-CRF[5]. 以中文字向量、词典特征作为输入，使用残差扩张卷积神经网络结合 CRF 层识别实体，在 CCKS17 数据集上取得了比主流 BiLSTM-CRF 模型更好的结果.

(4) CRF(Yang 和 Huang)[4]. 使用了字向量、部首、词性、拼音、词典和规则特征作为输入，CRF 模型识别实体，在 CCKS 2018 评测中获得了第一名.

(5) Ensemble(Luo 等人)[38]. 我们之前参加 CCKS 2018 评测的工作，集成了五个神经网络模型的结果，在评测中取得了第三名.

(6) BiLSTM-CRF(Ji 等人)[7]. 以字向量为输入，使用 BiLSTM-CRF 模型进行实体识别，然后构建了一个药物词典来修正药物实体，最后采用了一系列规则后处理来提升性能.

(7) Lattice LSTM. Zhang 等人[17]提出的 Lattice LSTM 模型，在基于字的中文 NER 上充分利用词

表 6　与其他方法的性能比较　　(单位：%)

方法	CCKS17						CCKS18					
	症状体征	检查检验	疾病诊断	治疗	身体部位	总体	解剖部位	症状描述	独立症状	药物	手术	总体
HIT-CNER[37]	96.00	94.43	78.97	81.47	87.48	91.14	—	—	—	—	—	—
BiLSTM-CRF-DIC[6]	—	—	—	—	—	91.24	—	—	—	—	—	—
RD-CNN-CRF[5]	—	—	—	—	—	91.32	—	—	—	—	—	—
CRF(Yang 和 Huang)[4]	—	—	—	—	—	—	87.97	90.59	**92.45**	**94.49**	85.43	89.13
Ensemble(Luo 等人)[38]	—	—	—	—	—	—	87.59	90.77	91.72	91.53	86.41	88.63
BiLSTM-CRF(Ji 等人)[7]	—	—	—	—	—	—	86.65	89.13	90.69	91.15	85.61	87.68
Lattice LSTM	**96.72**	93.87	78.69	77.16	86.20	90.51	87.58	90.76	91.99	84.21	86.14	88.09
BERT	95.56	93.11	77.84	80.17	86.01	89.94	88.17	89.08	90.52	84.33	84.66	88.00
CNN-CRF	95.16	93.97	76.57	78.54	87.38	90.36	86.23	88.78	90.80	91.63	80.35	87.17
CNN-CRF+ELMo	96.20	94.39	77.83	80.73	86.38	90.74	87.94	89.27	91.21	91.99	83.23	88.57
CNN-CRF-SP+ELMo	96.33	94.56	**82.07**	83.35	86.63	91.24	88.81	91.66	91.02	91.57	82.17	89.07
BiLSTM-CRF	95.94	93.99	76.55	81.63	87.19	90.69	87.17	89.46	91.20	91.29	85.04	88.08
BiLSTM-CRF+ELMo	95.13	94.75	80.82	**83.86**	88.19	91.51	88.44	91.08	91.63	91.91	**86.75**	89.28
BiLSTM-CRF-SP+ELMo	95.37	**94.94**	81.13	83.32	**88.74**	**91.75**	**89.69**	**91.83**	92.01	91.30	86.22	**90.05**

信息，并在通用领域上达到了先进水平. 为了进行公平的比较，我们使用他们论文提供的代码，以本文同样的生物医学预训练向量作为输入，在 CCKS 数据集上重新训练模型进行比较.

(8) BERT. Devlin 等人[16]提出的基于 Transformer 模型的预训练方法. 为了比较，我们使用官方提供的中文字 BERT 模型(https://github.com/google-research/bert)在 CCKS CNER 数据集上进行微调得到结果.

(9) CNN-CRF. 本文使用 w2v 字向量、词向量、词典和部首特征的 CNN-CRF 模型结果.

(10) CNN-CRF＋ELMo. 在上述 CNN-CRF 基础上加入 512 维笔画 ELMo 的结果.

(11) CNN-CRF-SP＋ELMo. 文本使用 w2v 字向量、词向量、词典、部首特征和 512 维笔画 ELMo 作为输入，基于私有共享多任务的 CNN-CRF 模型结果.

(12) BiLSTM-CRF. 本文使用 w2v 字向量、词向量、词典和部首特征的 BiLSTM-CRF 模型结果.

(13) BiLSTM-CRF＋ELMo. 在上述 BiLSTM-CRF 基础上加入 512 维笔画 ELMo 的结果.

(14) BiLSTM-CRF-SP＋ELMo. 文本使用 w2v 字向量、词向量、词典、部首特征和 512 维笔画 ELMo 作为输入，基于私有共享多任务的 BiLSTM-CRF 模型结果.

从结果可以看到，在数据集规模比较小的情况下，通过丰富的人工特征，CRF 模型同样能够取得先进结果. 相比传统的 BiLSTM，充分利用词信息的 Lattice LSTM 能够获得更好的效果. 通过多模型集成也能进一步提升结果. 相比其它方法，本文通过加入笔画 ELMo 在 BiLSTM-CRF 单模型下未使用任何后处理就能够获得当前先进结果. 从具体每类实体分析，在 CCKS18 上 Yang 等人的 CRF 方法在药物实体上有明显的优势. 主要原因是他们利用了多个药物相关网站来构建了更高质量的药物词典，也说明了我们方法通过更高质量的字典仍有上升空间. 相比 BERT 预训练方法，本文方法获得了更好的结果. 主要原因在于该 BERT 模型使用的是新闻领域语料进行预训练，而本文提出的 ELMo 则使用同领域生物医学语料进行预训练，且考虑了笔画特征，更能捕获汉字内部结构信息. 总体上看，在 CNN-CRF 和 BiLSTM-CRF 模型上，加入 ELMo 后，每类实体都获得了不同程度的提升，说明了本文 ELMo 向量的有效性. 在所有模型中，本文提出的 BiLSTM-CRF-SP＋ELMo 多任务模型在两个数据集上都取得了总体最好的 F 值，相比单任务的 BiLSTM-CRF＋ELMo 模型，多任务模型在 CCKS17 的身体部位和 CCKS18 的解剖部位、症状描述实体上有较为明显的提升. 主要原因是因为这几类实体在这两个数据集上相关性较强，通过多任务学习能够有效地学习到任务间的相关信息，提升了模型的性能.

4.7 ELMo 向量可视化分析

为了进一步分析笔画 ELMo 字向量学习到的表示，我们随机从数据集中抽取了一个句子：“取病理提示胃恶性肿瘤，给予奥沙利铂＋多西他赛方案化疗，近期患者精神可，大便次数多，小便正常，无腹痛、腹胀等不适.”进行传统 w2v 和笔画 ELMo 字向量可视化对比. 具体地，先获取该句中每个中文汉字的字向量表示，然后使用主成分分析(PCA)方法将其字向量降到 2 维，画出其散点图，如图 6 所示. 图中每个字旁的编号代表句子中的字索引号.

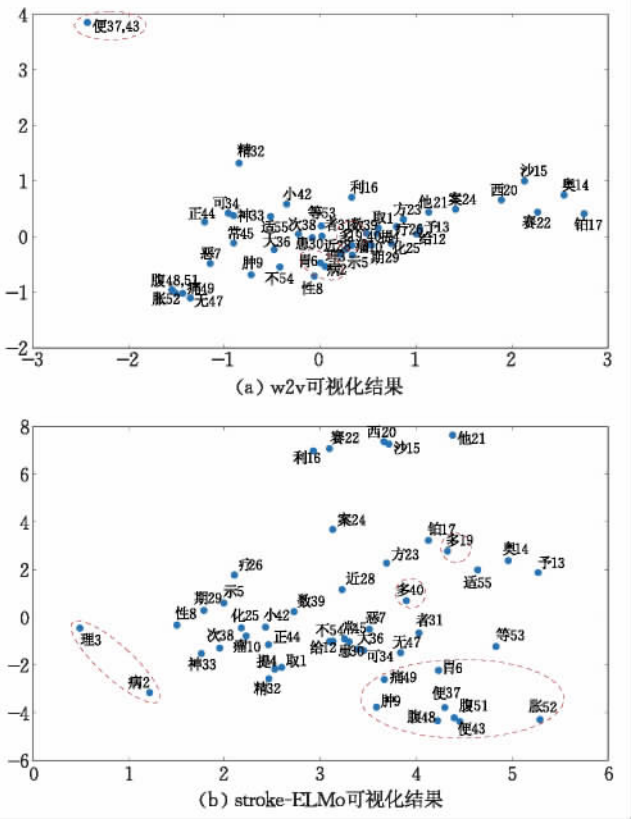

图 6 不同中文字向量可视化对比

从可视化结果可以看出，相比传统的 w2v 字向量，笔画 ELMo 可以对同一个字的不同语义进行不同的表示. 例如，句子中的“多”字，一个是药名“多西他赛”的“多 19”，一个是“大便次数多”的“多 40”. 可以看到传统的字向量对于“多”只能有一个字向量表

示，而在笔画 ELMo 中，不同的“多”有不同的向量表示. 且“多 19”在向量空间中更接近于“西、他、赛”几个字的向量；“多 40”更接近于“大、便”的字向量.

在 w2v 字向量空间中，我们可以观察到，一些在整个语料库中常出现的词语搭配的距离会比较近，例如“胃 6 病 2”；而“便”字离其它字比较远，可能是因为“便”字语义比较丰富，单独使用一个向量表示时，和本句的其它字较为不相似. 而在笔画 ELMo 向量空间中，则是具体上下文相关的字向量，即根据当前句子索引相近的字距离比较接近，例如“病 2”与“理 3”比较近，“胃 6”和“恶 7”比较近.

此外，还可以看出在笔画 ELMo 空间中，一些具有相似内部结构的字，其向量的距离较近. 例如，“腹”、“胀”、“胃”、“肿”等几个字都具有“月”字结构，基本都聚集在空间图的右下部. 这也说明了笔画 ELMo 根据笔画序列信息能够在一定程度上自动学习到字的内部结构信息.

5　总结与展望

本文提出了一种基于笔画 ELMo 和多任务学习的中文电子病历实体识别方法. 通过在大规模数据上语言模型预训练，笔画 ELMo 能够学习到丰富的汉字内部结构信息，且是上下文相关的字向量，优于传统字向量表示. 此外，利用多任务学习框架可以充分学习多个相关任务互补信息，进一步提升模型在各自任务上的性能. 在 CCKS17 和 CCKS18 两个 CNER 数据集上实验表明，本文提出的基于笔画 ELMo 和多任务学习的神经网络模型能够有效地识别中文电子病历实体. 此外，在 CNER 任务上，BiLSTM-CRF 模型优于 CNN、BiLSTM 和 CNN-CRF 模型. 通过加入其它额外特征的比较，其中加入高质量词典特征最为有效.

从 Lattice LSTM 模型结果可以看到，在基于字的中文 NER 模型上充分利用词信息具有很大潜力，如何设计模型更充分地利用字词信息来提升中文 CNER 效果，是我们下一步工作重点. 此外，本文提出的笔画 ELMo 是一种通用的中文向量表示方法，在我们未来的工作中，将探索其在更多中文 NLP 任务上的表现.

参　考　文　献

[1] Bethard S, Savova G, Chen W-T, et al. SemEval-2016 Task 12: Clinical TempEval//Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). San Diego, USA, 2016:1052-1062

[2] Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. Journal of the American Medical Informatics Association, 2013, 20(5): 806-813

[3] Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. Journal of the American Medical Informatics Association, 2010, 17(5): 514-518

[4] Yang X, Huang W. A conditional random fields approach to clinical name entity recognition//Proceedings of the Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing (CCKS 2018). Tianjin, China, 2018: 1-6

[5] Qiu J, Zhou Y, Wang Q, et al. Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field. IEEE Transactions on NanoBioscience, 2019, 18(3): 306-315

[6] Wang Q, Zhou Y, Ruan T, et al. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. Journal of Biomedical Informatics, 2019, 92(103133): 1-9

[7] Ji B, Li S, Yu J, et al. A BiLSTM-CRF method to Chinese electronic medical record named entity recognition//Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence. Sanya, China, 2018: 1-6

[8] Peters M, Neumann M, Iyyer M, et al. Deep contextualized word representations//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, USA, 2018: 2227-2237

[9] Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. Minneapolis, USA, 2019: 4171-4186

[10] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018-improving.pdf, 2018

[11] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch. The Journal of Machine Learning Research, 2011, 12(76): 2493-2537

[12] Strubell E, Verga P, Belanger D, et al. Fast and accurate entity recognition with iterated dilated convolutions//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 2670-2680

[13] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:150801991, 2015

[14] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition//Proceedings of the NAACL-HLT. San Diego, USA, 2016: 260-270

[15] Ma X, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany, 2016: 1064-1074

[16] Zhang Hai-Nan, Wu Da-Yong, Liu Yue, et al. Chinese named entity recognition based on deep neural network. Journal of Chinese Information Processing, 2017, 31(4): 28-35(in Chinese)
(张海楠, 伍大勇, 刘悦等. 基于深度神经网络的中文命名实体识别. 中文信息学报, 2017, 31(4): 28-35)

[17] Zhang Y, Yang J. Chinese NER using lattice LSTM//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, 2018: 1554-1564

[18] Zhao H, Kit C. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition//Proceedings of the 6th SIGHAN Workshop on Chinese Language Processing. Hyderabad, India, 2008: 106-111

[19] Peng N, Dredze M. Improving named entity recognition for Chinese social media with word segmentation representation learning//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 149-155

[20] Liu Bing-Yang, Wu Da-Yong, Liu Xin-Ran, et al. Chinese named entity recognition incorporating global word boundary features. Journal of Chinese Information Processing, 2017, 31(2): 86-91(in Chinese)
(刘冰洋, 伍大勇, 刘欣然等. 融合全局词语边界特征的中文命名实体识别方法. 中文信息学报, 2017, 31(2): 86-91)

[21] Meng Y, Li X, Sun X, et al. Is word segmentation necessary for deep learning of Chinese representations?//Proceedings of the 57th Conference of the Association for Computational Linguistics. Florence, Italy, 2019: 3242-3252

[22] Li H, Hagiwara M, Li Q, et al. Comparison of the impact of word segmentation on name tagging for Chinese and Japanese//Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland, 2014: 2532-2536

[23] Harris Z S. Distributional structure. Word, 1954, 10(2-3): 146-162

[24] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality//Proceedings of the Advances in Neural Information Processing Systems. San Diego, USA, 2013: 3111-3119

[25] Pennington J, Socher R, Manning C D. GloVe: Global vectors for word representation//Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014). Doha, Qatar, 2014: 1532-1543

[26] Yang L, Sun M. Improved learning of Chinese word embeddings with semantic knowledge//Proceedings of the Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Guangzhou, China, 2015: 15-25

[27] Xu J, Liu J, Zhang L, et al. Improve Chinese word embeddings by exploiting internal structure//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA, 2016: 1041-1050

[28] Chen X, Xu L, Liu Z, et al. Joint learning of character and word embeddings//Proceedings of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 1236-1242

[29] Yu J, Jian X, Xin H, et al. Joint embeddings of Chinese words, characters, and fine-grained subcharacter components//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 286-291

[30] Cao S, Lu W, Zhou J, et al. cw2vec: Learning Chinese word embeddings with stroke *n*-gram information//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 5053-5061

[31] Liu P, Qiu X, Huang X. Adversarial multi-task learning for text classification//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada, 2017: 1-10

[32] Zhao S, Liu T, Zhao S, et al. A neural multi-task learning framework to jointly model medical named entity recognition and normalization//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019: 817-824

[33] Crichton G, Pyysalo S, Chiu B, et al. A neural network multi-task learning approach to biomedical named entity recognition. BMC Bioinformatics, 2017, 18(1): 368

[34] Che W, Liu Y, Wang Y, et al. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and Treebank concatenation//Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Brussels, Belgium, 2018: 55-64

[35] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. Journal of Machine Learning Research, 2012, 13(2): 281-305

[36] Prechelt L. Automatic early stopping using cross validation: quantifying the criteria. Neural Networks, 1998, 11(4): 761-767

[37] Hu J, Shi X, Liu Z, et al. HITSZ_CNER: A hybrid system for entity recognition from Chinese clinical text//Proceedings of the Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing (CCKS 2017). Chengdu, China, 2017: 1-6

[38] Luo L, Li N, Li S, et al. DUTIR at the CCKS-2018 Task1: A neural network ensemble approach for Chinese clinical named entity recognition//Proceedings of the Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing (CCKS 2018). Tianjin, China, 2018: 1-6

LUO Ling, Ph. D. candidate. His main research interests include biomedical text mining, machine learning and natural language processing.

YANG Zhi-Hao, Ph. D., professor. His main research interests include text mining, machine learning and natural language processing.

SONG Ya-Wen, M. S. candidate. Her main research interest is biomedical text mining.

LI Nan, Ph. D. candidate. Her main research interests include biomedical text mining and knowledge graph.

LIN Hong-Fei, Ph. D., professor. His main research interests include search engine, text mining, sentiment analysis and natural language processing.

Background

The task we have studied in this paper is Chinese clinical named entity recognition (CNER) which is in the area of biomedical text mining. The Chinese CNER task aims to identify and extract the related medical clinical entities (e. g., anatomy, symptom, independent symptom, drug and operation) from Chinese clinical text. Most of these existing Chinese NER works often follow the English processing methods. In addition, the existing Chinese CNER dataset size is small, and medical entity annotation requires medical background knowledge, which is time-consuming and labor-intensive. In this paper, we propose a Chinese CNER method based on stroke ELMo and multi-task learning. Moreover, the effects of common NER features and neural network models were investigated for the Chinese clinical entity recognition task. The experimental results show that the BiLSTM-CRF model outperforms the CNN-CRF model. When our stroke ELMo is added, the BiLSTM-CRF model can achieve an average improvement of 1.01% in *F*-score on both datasets. Moreover, the additional NER features can further improve the performance. Our neural network model based on stroke ELMo and multi-task learning achieves the state-of-the-art results on the CCKS17 and CCKS18 CNER datasets (the *F*-scores of 91.75% and 90.05%, respectively).

This work is supported by the National Key Research and Development Program of China (No. 2016YFC0901900). It focuses on Chinese CNER, which is a part of key technologies and systems for medical information extraction. This research project aims at building a precision medicine knowledgebase application platform, which includes the information from biological molecular, human disease, phenotype, drugs, etc. And it will support functions of retrieval by various biomedical terms, workflow analyses for omics datasets and the intelligent knowledge discovery.