这篇文章展示了一种中文医疗实体识别方法，结合了中文笔画elmo和 迁移学习。首先，使用了一些预处理方法包括：句子分割，生成笔画等。第二步，使用cw2vec工具在大量的无标注数据上学习了传统的字embedding向量。第三步，将中文笔画elmo引入到模型中。第四步，将embedding作为输入，BILSTM-crf在标注好的训练集上训练。为了充分使用已有的相关数据，我们使用CCKS2018的语料用做迁移学习。第5步，我们使用不同的随机种子训练了5个模型，而且使用多数人投票将这些模型结合起来。另外一些还采用了一些后处理的方法。我们模型的细节将在后续章节介绍。

特征

传统字embedding向量由于在中文NER中字的特征比词的特征更优秀，因此我们的方法中使用了字符的特

征。为了获得更高质量的字embedding，我们下载了中文CNKI的医疗摘要然后收集了由CCKS提供的中文电子医疗记录信息，总共有156w条文本，作为无标签数据。然后这些文本被用于训练200维的字embedding向量作为预训练的传统字embedding，使用cw2vec工具完成训练。

笔画elmo

上面提及的传统字嵌入方法仅仅学习了一个字符的全局向量表示。然而，一个字符在特定上下文中可能不同的含义。例如，根据上下文的含义，多的意思可能会不同。很合理的认为，字符“多”基于两种不同的单词含义，应该有两种不同的向量表示。近期，elmo（从语言模型中学习词向量）就是这么一种能提供深度上下文词向量的方法。传统的elmo基于英文文本提出的，而且它对一个单词用到的所有上下文都生成了词向量，因为对相同的单词允许有不同的向量表示。che et al将elmo应用到更多的语言，包括中文 。

和这种方法不同，我们提出了笔画elmo模型来学习语境化的中文“字”向量。图二展示这种笔画elmo的结构。受到英文字向量的启发，我们使用中文笔画序列信息来构造笔画elmo模型。我们想法是有相关中文语义的字符通常有相似的内部结构。例如，“森”，“林”，“木”都是由不同数量的“木”组成。因此，中文字符的基础笔画序列被用作语言模型的输入。然后，在大规模语料上预训练语言模型来学习内部的联系。具体来讲，首先在语料词汇表中每个字的笔画序列通过handian网站获得。然后，笔画的embedding向量被输入到卷积层。使用max pooling层来抽取卷积层中的特征。接着，这些特征向量被输入到基于BILSTM的双向语言模型中。在使用上述提及的无标注数据训练双向语言模型后，笔画elmo抽 取biLM的中间层表示，然后接上全连接获取512维的语境字embedding。

bilstm-crf 模型

和很多ner任务一样，我们对cner建模成一个序列标注问题。使用BIOES标注方法。我们采用了BILSTM-CRF模型作为CNER任务，这种结构在图3中有介绍。第一步：一个句子展示成embedding向量序列。第二步，将这个embedding作为BILSTM层的输入。双向LSTM有两个lstm分别获取从左到右的特征和从右到左的特征。第三步，对bilstm的输出使用tanh激活函数获得更高维的特征。最后使用CRF在所有可能的标注路径中获得最优的标注序列路径。

迁移学习

在这节中我们介绍了我们模型中迁移学习的细节。为了更好的利用已有数据，在CCKS2018语料上使用迁移学习进一步提高模型表现。 CCKS2018数据是一个CNER任务，但是抽取实体类型和CCKS2019任务并不相同。 CCKS2018需要确定和抽取相应的医疗实体从中文医疗数据中。尽管标注规则是不一样的，但是它们在结构上是相关的，药品和手术。所以我们使用了CCKS2018 CNER作为原领域，然后CCKS2019 CNER作为目标领域来迁移学习。在深度学习和迁移学习方面指定一个神经网络层， 然后共享部分或者所有的参数。我们使用三种不同结构来做CNER，如图4.

全连接迁移学习模型（FS-TL）.

在FS-TLL模型中，除了输出层所有模型的参数都共享 。每个任务都有一个任务指 定的CRF层，用于最终tanh层做预测。

分成迁移学习模型。

在H-TL模型中，embedding层和最初的BILSTM层的参数都被共享。每个任务都有个任务特定的tanh layer和crf输出层，CCKS 2019任务额外添加了一个任务特定的BILSTM层。 私有分享迁移学习模型(sp- tl).在sp-tl模型中，除了共享embedding，bilstm和tanh层外，每个人物都有一个任务指定的BILSTM和tanh层。 然后共享部分和特定任务部分被拼接在一起，输入到任务特定的CRF输出层中。这种。结构使得模型能够选择性的利用共享或者指定任务的信息。

2.4 集成和后处理

为了缓解随机初始化导致的模型影响，我们使用不同是随机种子训练了5种模型，然后使用多数表决的方式来结合预测结果。另外，为了使结果最优，我们还采用了几种普遍的后处理方法，例如tagging 均衡，删除空格和括号平衡。