基于 CRF 和 Bi-LSTM 的命名实体识别

1160300607 张开颜

哈尔滨工业大学计算机学院

minekaiyan@gmail.com

摘要

命名实体识别任务主要是识别文本中的人名、地名、机构名等实体。基于 CRF 的命名实体识别是传统的识别方法，而随着深度神经网络的发展，深度学习也逐渐应用到实体识别中。本文将首先介绍命名实体识别的相关工作，然后介绍 CRF 模型原理，以及用于命名实体识别的深度神经网络 Bi-LSTM 网络。之后实现基于 CRF 和 CRF+Bi-LSTM 的两种命名实体识别模型，并使用 BioNLP 数据集进行实验分析。实验中主要分析了 CRF 模型的特征选取和组合对模型的影响，同时进行了参数调整以获取最优 CRF 实体识别模型。之后分析 CRF+Bi-LSTM 模型各参数对模型性能的影响并做相关调整以获得最优性能，最后将 CRF 模型和 CRF+Bi-LSTM 模型的命名实体识别效果进行对比。实验结果表明，CRF+Bi-LSTM 模型相对于 CRF 模型有更好的识别效果。

关键字：CRF；Bi-LSTM；命名实体识别

Named Entity Recognize Based On CRF and Bi-LSTM

1160300607 Zhang Kaiyan

Harbin Institute of Technology

minekaiyan@gmail.com

Abstract

The named entity recognition task mainly identifies entities such as person names, place names, and institution names in the text. CRF-based named entity recognition is a traditional recognition method, and with the development of deep neural networks, deep learning is gradually applied to entity recognition. This article will first introduce the current state of named entity recognition, and introduce the CRF model, as well as the deep neural network Bi-LSTM network for named entity recognition.Two named entity recognition models based on CRF and CRF+Bi-LSTM were then implemented and experimental analysis was performed based on the BioNLP data set. It mainly

analyzes the influence of feature selection and combination of CRF model on the model, and adjusts the parameters to obtain the optimal CRF entity recognition model. After that, the influence of each parameter of CRF+Bi-LSTM model on the performance of the model is analyzed to obtain the optimal performance. Finally, the CRF model and CRF+Bi-LSTM model are compared for the performance of named entity recognition. The experimental results show that the CRF+Bi-LSTM model has a better recognition effect than the CRF model.

Keywords：CRF；Bi-LSTM；NER

0 引言

命名实体识别（Named Entity Recognition, NER）在 MUC-6 第一次被提出[1]，现在在自然语言处理中已经被广泛应用。其目的是识别出文本中表示命名实体的成分，并对其进行分类。命名实体识别技术是信息抽取、信息检索、句法分析、机器翻译、问答系统等多种自然语言处理技术必不可少的组成部分。从语言分析的角度来看，命名实体具有随意性、复杂性以及多变性等特点，命名实体识别属于词法分析中未登录词识别的范畴，至今仍然是一个重要且具有挑战性的研究课题。

早期的命名实体识别工作，主要识别一般的专有名词[2]，包括三类名词：人名、地名、机构名。随着研究范围的扩大，针对不同的特定问题特定领域，越来越多的实体类型被提出。在生物医学领域，对于基因名、蛋白质名的识别已经有许多工作在开展，也取得了不错的效果。在本文中将针对生物学领域方面，在 BioNLP/NLPBA 2004 的的 GENIA 数据集[3] 进行命名实体识别相关技术的实现。

1 相关工作

命名实体识别研究至今已经有近二十年的发展历史，已经成为了自然语言处理领域的一项重要技术，根据模型和算法的不同，先已陆续推出了成效可观的各类技术成果。

规则和词典相结合的方法最早应用于命名实体识别中，其多采用语言学专家手工构造规则模板，选用特征包括标点符号、关键字等方法，以模式和字符串匹配为主要手段。采取这种方法的代表性系统包括 GATE 项目中的 ANN E 系统以及参加 MUC 评测的 FACLE 系统，这类系统大多依赖于知识库和词典的建立，缺点是代价太大，存在系统建设周期长、移植性差等问题。如王宁[4] 等利用规则的方法进行金融领域的公司名识别，该系统对知识库的依赖性强，同时开放和封闭测试的结果也显示了规则方法的局限性。

传统的命名实体识别模型往往是基于统计的方法，在 CoNLL-2003 会议上，所参赛的 16 个系统全部采用基于统计的方法。基于统计的方法利用人工标注的语料进行训练，标注语料时不需要广博的语言学知识，并且可以在较短时间内完成。基于统计机器学习的方法主要有：

隐马尔可夫模型（HMM）、最大熵模型（ME）、最大熵马尔可夫模型（MEMM）、支持向量机（SVM）以及条件随机场（CRF）等。Zhao 等通过最大熵模型对 4 类名词进行实体识别，获得 77.87% 的准确率 [5]。另外有陈霄 [6] 采用 SVM 模型提出了中文组织机构名的实体识别，准确率达到了 81.68%。其中 CRF 自 2001 年由 Lafferty 等人 [7] 提出后，就广泛应用于命名实体识别领域，如 Settles [8] 等使用 CRF 对生物医学命名实体进行识别；姜文至等 [9] 提出的基于 CRF 和规则相结合的军事命名实体识别模型。

近年来，随着深度学习的不断发展，深度神经网络被广泛适用于命名实体识别模型中。深度神经网络通过搭建多层网络结构实现数据的深加工，目前比较常用的结构包括卷积神经网络（CNN）、循环神经网络（RNN）以及长短时记忆（LSTM）等。多数基于深度神经网络的方法重视整体上的全局特征，认为文本中的每一个字都对其他字的判断产生影响，而忽视了文本具有的局部特征。因此，如何在特征提取过程中加入文本局部特征成为提高命名实体效果的关键。

鉴于传统统计模型和深度神经网络的特点，将二者结合是如今研究的主流趋势。Huang 等 [10] 首次将双向 LSTM 和 CRF 模型相结合，并运用于序列标注任务中，显著改善了命名实体识别的效果。Liu 等 [11] 在此基础上对序列标注任务进行加强，提出了 LM-LSTM-CRF 模型。除了 LSTM 之外，卷积神经网络也被广泛用于命名实体识别模型中，如 Ma [12] 等提出的基于字和词输入特征卷积融合的 LSTM-CNNs-CRF 模型，同样取得了较好的效果。

2 基于 CRF 的命名实体识别

2.1 CRF

条件随机场（CRF）是一种用来标记和切分序列化数据的统计模型，可以把它看成一个无向图模型或马尔可夫随机场。当输入节点值给定时，可以用于计算指定输出节点值的条件概率。假设观察序列（需要标注的句子）为 $x = \{x_1, x_2, \cdots, x_n\}$，有限状态集合（标注序列）为 $y = \{y_1, y_2, \cdots, y_n\}$。根据条件随机场理论 [13] 有：

$$p(y|x, \lambda) \propto \exp\left(\sum_i \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k(y_i, x, i)\right) \tag{1}$$

其中 $t_j(y_{i-1}, y_i, x, i)$ 和 $s_k(y_i, x, i)$ 统称为特征函数。前者表示整个观察序列和相应的标记序列在 $i-1$ 和 i 时刻的特征，是一个转移函数；后者是 i 时刻整个观察序列和标记的特征，是一个状态函数。可以把两个特征函数统一为：

$$f_j(y_{j-1}, y_j, x, i) = \begin{cases} 1 & if\ y_{j-1}, y_j, x_i\ s.t \\ 0 & otherwise \end{cases} \tag{2}$$

由此可以得到 CRF 的概率公式形式：

$$p(y|x,\lambda) = \frac{1}{Z(x)} \exp(\sum_{i=1}^{n} \sum_{j} \lambda_j f_j(y_{i=1}, y_i, x_i)) \tag{3}$$

$$\text{其中} Z(x) = \sum_{j} \exp\left(\sum_{i=1}^{n} \sum_{j} \lambda_j f_j(y_{i-1}, y_i, x, i)\right) \tag{4}$$

则 y 的最大概率标记序列为 $y^* = \arg\max p(y|x,\lambda)$。

一般 CRF 模型运行主要分为三个步骤：

(1) 预定义特征函数 f

(2) 在给定的数据上，训练模型，确定参数 λ

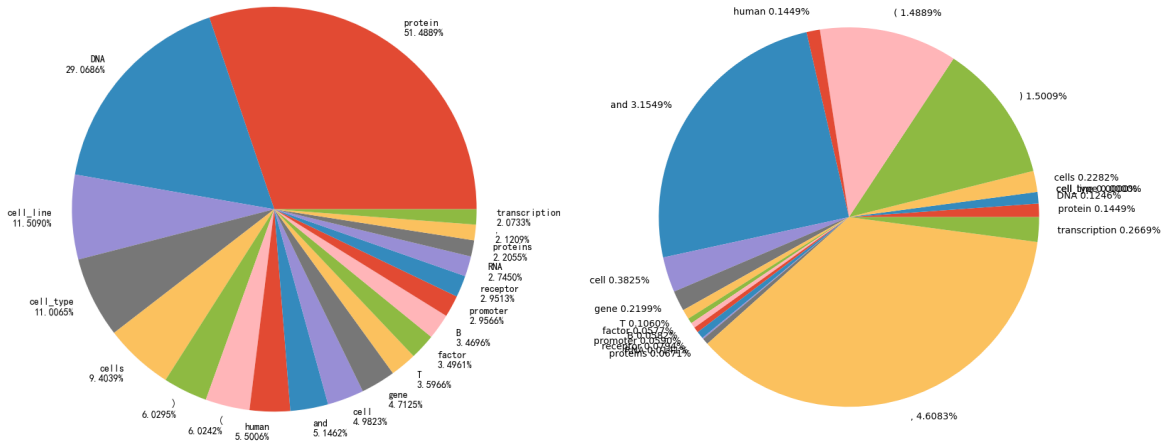
(3) 用确定的模型做序列标注问题或序列求概率问题

2.2 特征提取

特征提取是 CRF 模型中非常重要的一部分，如何针对特定的任务为模型选取有效的特征，对于模型性能有很大的影响。在命名实体识别中，常用的特征类型包括：当前词、上下文、词的子串、词型（包含大小写、数字、符号等）、词的标签以及上下文标签（词性标签）等。

实验给定 BioNLP/NLPBA 2004 的 GENIA 数据集，主要包括 DNA、RNA、细胞系（cell line）和细胞类型（cell type）四种命名实体。通过提取训练集中所有命名实体，统计分析命名实体的构成规律：

(1) 命名实体中存在常用词。如图1(a)，给出了命名实体中前 20 的高频词，其中 protein（蛋白质）在命名实体中出现频率高达 51.4889%，而 DNA 也在 29.0686% 的命名实体中出现。对比分析，如图1(b)，在非命名实体中，protein 和 DNA 仅占比 0.1%。由此说明有些词在命名实体中较为常见，而在非命名实体中较少出现。因此可以把当前词作为命名实体的特征，同时也可以把上下的词作为特征。



(a) 命名实体中前 20 高频词　　(b) 前 20 高频词在非命名实体中占比

图 1: 命名实体和非命名实体高频词对比

(2) 命名实体通过存在通用的前缀和后缀，如图2，给出了命名实体中前缀和后缀（有长度为 2 和 3 两个情况）频次最高的 5 个在命名实体和非命名实体中的占比。可以发现大多数前缀和后缀，在命名实体和非命名实体中出现的频次具有较大差异。比如，-DNA、-ein、pro-、ce-等在命名实体中较为常出现，而在非命名实体中出现的较少；而-es、an-、re-等在非命名实体中出现的频率远大于在命名实体中出现的频率。因此可以在一定程度上通过前缀和后缀来识别命名实体和非命名实体。

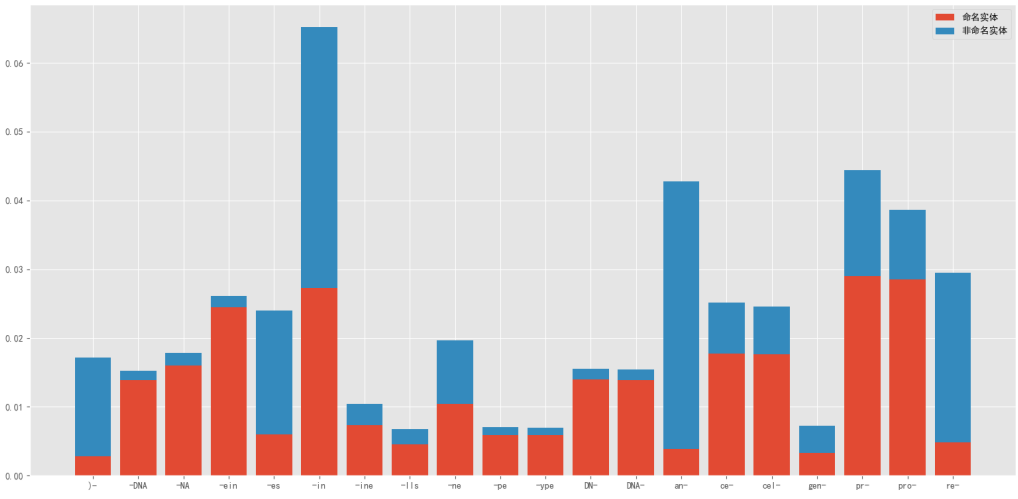


图 2: 命名实体和非命名实体前后缀占比

(3) 通过前述分析，命名实体通常满足一定的规则。如图3给出了几种正则表达式分别在

命名实体和非命名实体中所占比例。可以全大写的单词（isupper）、包含数字和'-' 的单词有更大的概率是命名实体。而全是数字的字符串中非命名实体占比更多。

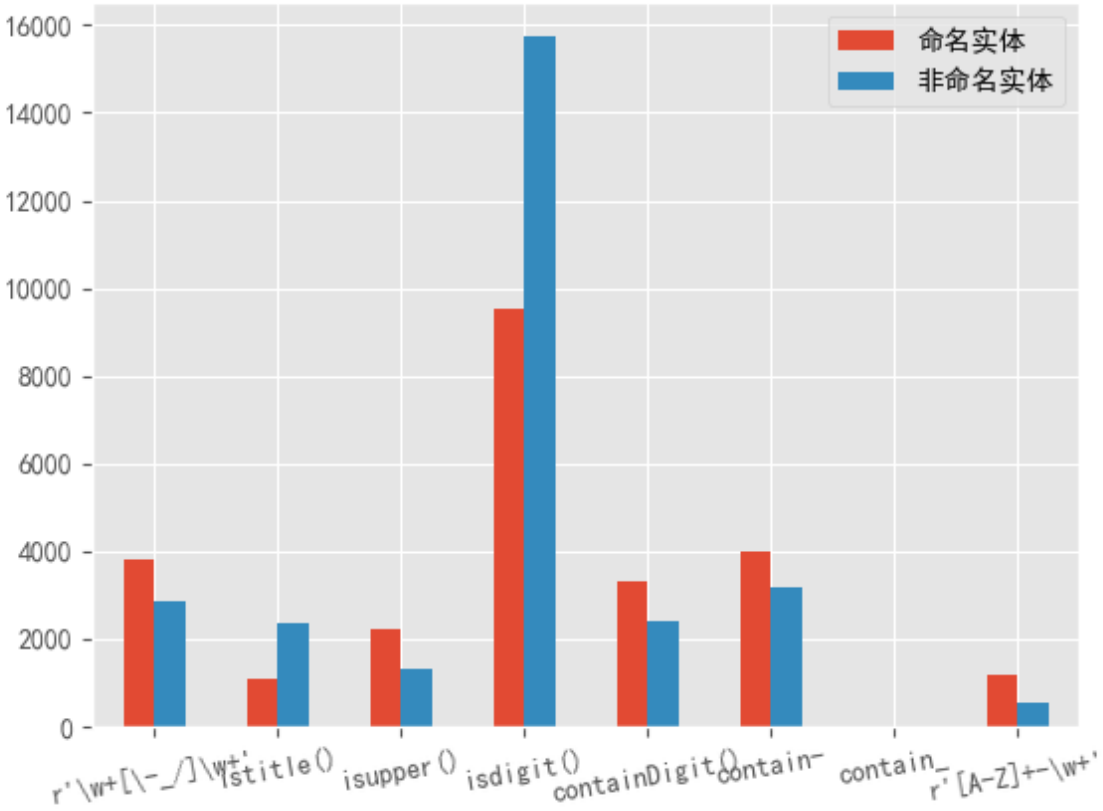


图 3: 命名实体和非命名实体构成正则

(4) Nothman 等[14] 在使用 Wiki 语料库进行实体识别训练时，发现词性特征在实体识别研究中具有较为重要的作用。因此，选择使用 NLTK 对需要识别的句子进行词性标注，然后将词性标注结果作为实体特征进行训练。

综上，给出如表1中对当前词预提取的特征及说明，其中使用'\w+[_/]\w+'（正则 1）和'[A-Z]+-\w+'（正则 2）作为命名实体的匹配正则。考虑到上下文的影响，同时对当前词的前 2 个词和后 2 个词提取如表1的特征作为当前词的特征。

表 1: 预提取的所有特征及相关说明

当前词特征	特征说明
word	当前词
word.isupper	当前词是否大写
word.istitle	当前词首字母是否大写
word.isdigit	当前词是否是数字
word.contain_digit	当前词是否包含数字
word.contain_signal	当前词是否包含-、/、_ 等符号
word[-3:]	当前词的后三个字母
word[-2:]	当前词的后两个字母
word[:2]	当前词的前 2 个字母
word[:3]	当前词的前 3 个字母
word.regular	是否满足正则
word.tag	当前词词性
+1:word	当前词后一个词
-1:word	当前词前一个词

2.3 CRF 模型框架

实现的 CRF 模型流程如图4，首先加载数据集，以句子为单位保存，其中句子保存为 xseq，对应的命名实体标记表示为 yseq。将每个句子作为一个样本，对句子中的单词提取特征。将所有的句子和对应的标记添加到模型中，然后设置模型参数进行训练。训练完成后，使用模型进行标记，得到句子的实体标记预测输出。

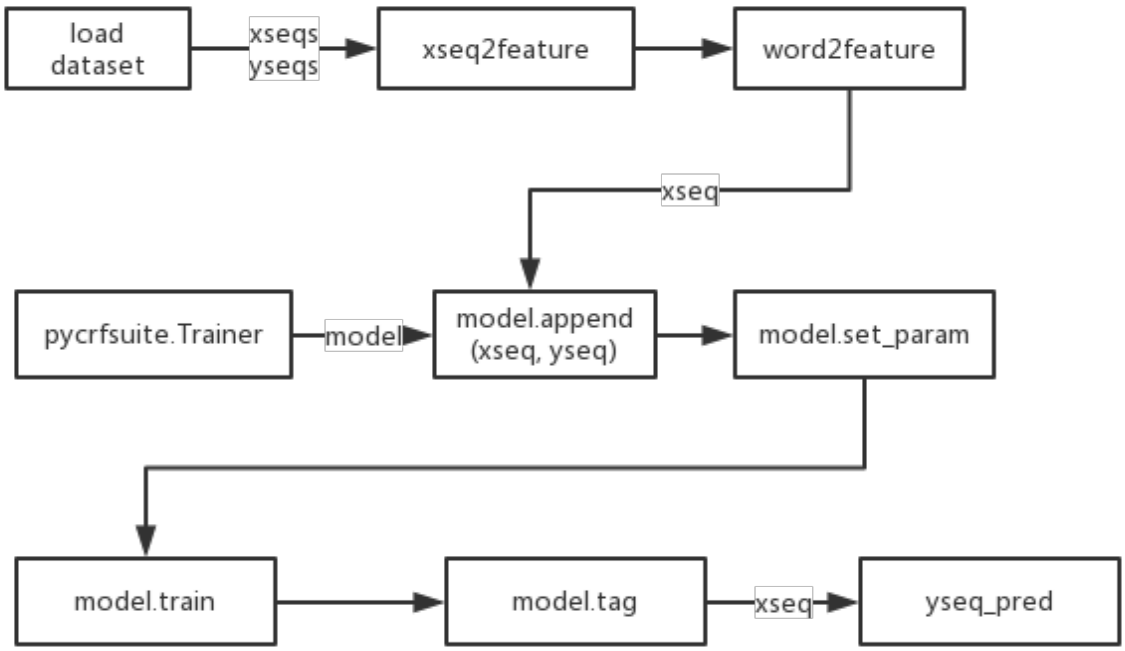


图 4: CRF 模型流程图

3 基于 CRF 和 LSTM 的命名实体识别

3.1 Bi-LSTM

长短时记忆网络 LSTM [15] 是一种解决序列标注中出现的长依赖问题的循环神经网络（RNN）模型。LSTM 相比于一般的 RNN 有着更为复杂的记忆装置，能够捕捉到序列中的长距离依赖。一般 LSTM 包含 3 个门，分别是：输入门（Input Gate）、忘记门（Forget Gate）、输出门（Output Gate），通过 3 个门来控制细胞状态。其中输入门决定保留当前输入的多少信息，忘记门决定保留上一个隐层传来的多少信息，输出门决定将输出多少信息。每个门通过 sigmoid 层和 pointwise 层的操作来对输入到门的信息进行选择和删除 [16]。例如 sigmoid 层通过产生一个 0-1 之间的参数来用来选择相应比例的信息。

Bi-LSTM 相较于 LSTM 又引入了一定程度上的优化。Bi-LSTM 包含前向和后向的两个 LSTM 序列网络，不仅可以保存前面的信息，同时也可以考虑到之后的信息。如图5，可以发现 Bi-LSTM 的隐藏层要保存两个值，A 参与正向计算，A' 参与发现计算，最终输出值 y 取决于 A 和 A'。正向计算时，隐藏层的 S_t 与 S_{t-1} 有关；反向计算时，隐藏层的 S_t 与 S_{t+1} 有关。

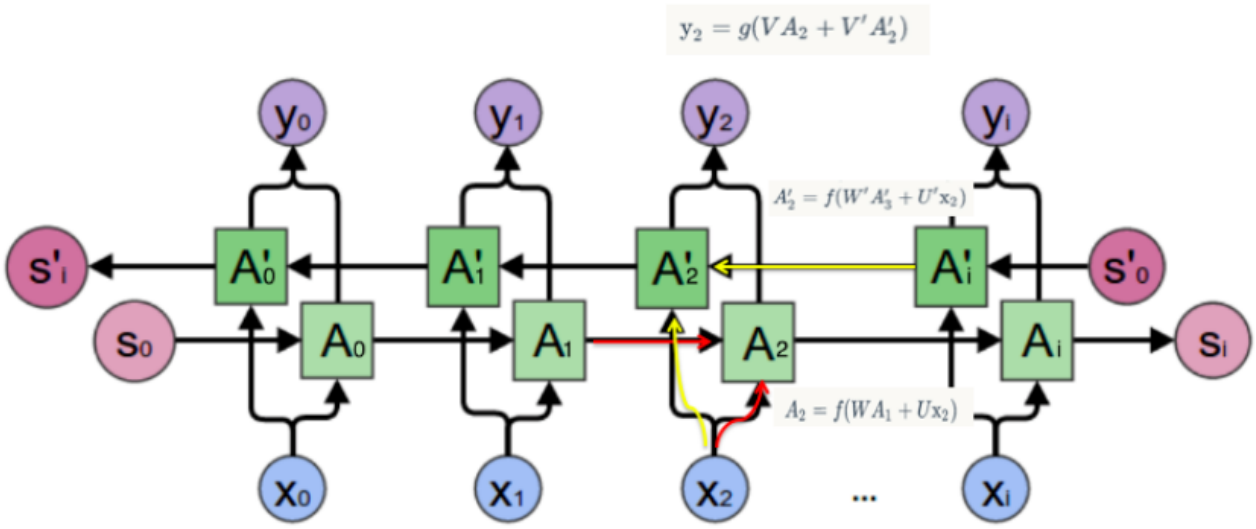


图 5: Bi-LSTM 网络结构

3.2 CRF+Bi-LSTM 模型框架

通过结合 Bi-LSTM 序列表示和 CRF 的序列标注，实验中搭建如图6的 CRF-BiLST 神经网络模型。

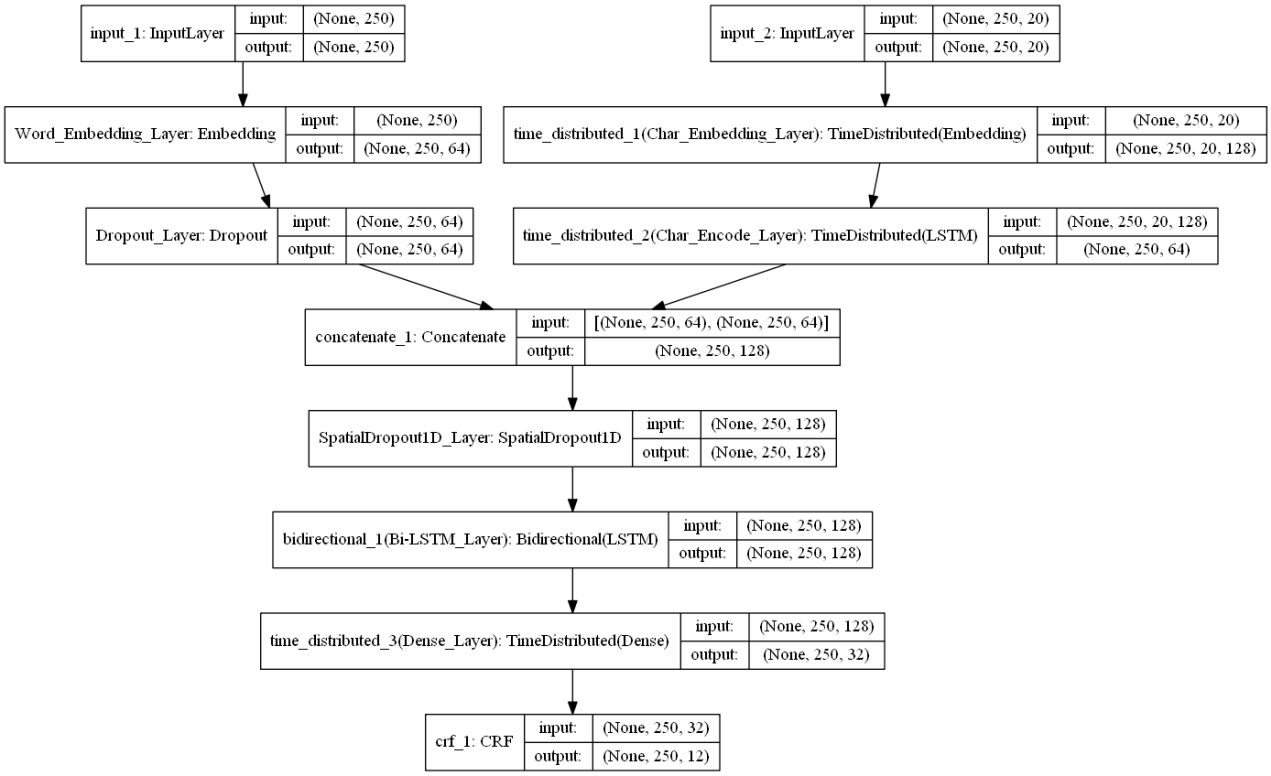


图 6: CRF+Bi-LSTM 网络模型

输入层（Input Layer）：对于输入的句子，首先构建单词表和字符表，然后将每个词用单词表中的序号表示，作为词向量表示输入；将每个词中每个字符用字符表中的序号表示，得到每个词的字符级别字向量，作为字符向量表示表示层输入。

词向量嵌入层（Word Embedding Layer）：通过 embedding 层，对输入的词转换成整数向量形式。本质是对输入词进行向量化升维，以获得词与词之间的联系以及更多的特征。为了防止过拟合，将嵌入结果之后加入 Dropout 层，在训练中每次更新时，将输入单元的按比率随机设置为 0。

字符向量嵌入层（Char Embedding Layer）：通过 embedding 层对输入的字符向量进行向量化升维。

字符向量编码层（Char Encoding Layer）：使用双向 LSTM 网络对字符向量进行训练，以学习到词的内部字符级别的组成特征，比如大小写等特征，然后进行一定的特征压缩降维。

拼接层（Concatenate Layer）：将词向量和字符向量拼接，作为整体输入到之后的双向 LSTM 网络中。之后通过 SpatialDropout1D 将有助于提高特征图之间的独立性，避免过拟合。

Bi-LSTM 层：将词向量和字符级别字向量进行拼接，整个作为特征输入到 Bi-LSTM，通过 Bi-LSTM 的长短时记忆的特征，对输入进行学习。

CRF 层：由于 LSTM 具有全局学习的特点，存在一定的不足，最后通过 CRF 层，结合双向 LSTM 输出层的序列以及最终给定的标注序列训练输出序列关于输入序列的条件概率模型。在判决时，给定双向 LSTM 隐藏层的输出向量序列，得到每个字属于每一个命名实体标签的概率向量。

4 实验

4.1 实验环境

实验软件环境如表2：

表 2: 实验所使用的软件及相关库

名称	说明
Python3.6	编程语言
Pycharm Professional	Python IDE
python-crfsuite [17]	CRF 模型实现库
keras2.2.4	CRF+BiLSTM 神经网络框架
tensorfow1.12	keras 框架后端
keras-contrib [18]	CRF+BiLSTM 的 CRF 层实现
nltk [19]	用于词性标注

实验硬件环境如图3：

表 3: 实验硬件环境

硬件名	配置	说明
笔记本	Windows 10 家庭中文版 (64 位) CPU (英特尔)Intel(R) Core(TM) i5-6300HQ CPU @ 2.30GHz 内存 8.00 GB	CRF 模型运行环境
服务器	GPU：gtx 1080ti(1 块) CPU：i5-7500 @ 3.40GHz(2 核) 内存：13G	CRF+LSTM 模型运行环境

4.2 实验数据

实验使用 BioNLP/NLPBA 2004 [20] 的数据集，来自于 GENIA 版本 3.02 语料库。并进行了简化，仅使用 protein（蛋白质）, DNA, RNA, cell line（细胞系）和 cell type（细胞类型）等命名实体。为了看出版年的影响，测试集大致分为四个子集：1978-1989 集（比训练集更早期），1990-1999 集（与训练集同时期），2000-2001 集（比训练集更晚期）和 S / 1998-2001 集（大致代表比训练集更晚期）。下表显示了数据集的大小:

表 4: 数据集

		abs	sentence	tokens
Training Set		2000	20546(10.27/abs)	472006(236.00/abs)(22.97/sen)
Test Set	Total	404	4260(10.54/abs)	96780(239.55/abs)(22.72/sen)
	1978-1989	104	991(9.53/abs)	22320(214.62/abs)(22.52/sen)
	1990-1999	106	1115(10.52/abs)	25080(236.60/abs)(22.49/sen)
	2000-2001	130	1452(11.17/abs)	33380(256.77/abs)(22.99/sen)
	S/1998-2001	204	2254(11.05/abs)	51628(253.08/abs)(22.91/sen)

实验时没有考虑时间边界问题，使用 Genia4ERtraining 文件夹下的 Genia4ERtask1.iob2 作为训练集;使用 Genia4ERtest 文件夹下的 Genia4EReval2.iob2 作为测试集;使用 JNLPBA2004_eval 文件夹下的 SharedTaskEval.pl 进行评价。

4.3 评价指标

本实验采用准确率 P(Precision)、召回率 (Recall) 以及 F1 值对模型的性能进行评价。3 个评价指标的定义如下：

$$P = \frac{\text{正确识别的命名实体个数}}{\text{识别的命名实体总数}} \tag{5}$$

$$R = \frac{\text{正确识别的命名实体个数}}{\text{命名实体总数}} \tag{6}$$

$$F1 = \frac{2 \times P \times R}{(P + R)} \tag{7}$$

4.4 实验结果及分析

4.4.1 CRF 模型

对于 CRF 模型而言，特征的选取十分重要。实验中，以之前预选取的特征作为基础，首先通过在 baseline 系统上进行系列单一特征增性实验，考察每个特征在 baseline 上的作用效果。然后选取特征作为整体加入模型，在此基础上进行系列单一特征减性实验，考察是否存在冗余特征，便于进一步优化。最后选取最优的特征进行模型调参，以获得最优模型。详细实验结果及分析如下：

(1) 使用当前词本身（包含本身及小写形式）为特征作为 CRF 模型的 baseline 系统，固定超参数和迭代次数（100 次）。每组实验都在 baseline 基础上加入一个特征，观察单一

特征对系统性能的影响程度，实验结果如表5，给出了在各时期的测试集上召回率、准确率和 F1 值的平均值。其中上下文 1 表示'+/-1:word'，上下文 2 表示'+/-2:word'，前后缀表示'word[:x]' 和'word[-x:]'。

表 5: 各特征在 baseline 的 CRF 模型上的作用效果

编号特征	R(%)	P(%)	F1(%)
当前词	57.20	64.25	60.46
+ 上下文 1	**61.84**	**69.27**	**65.29**
+ 上下文 2	**61.77**	**69.21**	**65.23**
+ 前后缀	**62.27**	**64.42**	**63.25**
+ 词形 1(istitle/upper/digit)	57.13	63.41	60.07
+ 词形 2(包含符号)	56.58	**64.31**	60.10
+ 正则 1	56.56	63.90	59.98
+ 正则 2	56.66	**64.30**	60.19
+ 词性	**57.98**	**64.65**	**61.08**
+ 所有增性特征	68.95	68.96	68.87
+ 所有特征	**70.01**	**69.76**	**69.80**

通过分析表5，可以发现上下文（当前词的前一个词和后一个词）的增性效果最好，同时发现当扩大上下文的窗口时（使用当前词的前两个和后两个词）作为特征时，效果反而下降，主要是因为此时发生了过拟合。其次增性效果较好的是当前词的前后缀，这比较符合之前的命名实体构成分析，主要是因为命名实体通常是名词，由此加-es 的词时命名实体的可能性较大。让人出乎意料的是正则表达式的效果反而较差，而词性特征的增性效果也不太理想。

最后，可以发现将所有增性特征组合的效果却没有将所有特征组合的效果好，这一不符合逻辑的现象说明特征组合存在不稳定性，其内部的原理仍然有待学习。

(2) 通过在 (1) 中的实验，可以发现将有增性和无增性的特征组合的效果反而比只组合有增性的特征的效果好。因此下面将在所有特征的组合上进行减性实验，考察是否存在冗余特征。实验结果如表6

表 6: 对 (1) 中所有特征组合模型进行减性实验

特征	R(%)	P(%)	F1(%)
所有特征	**70.01**	**69.76**	**69.80**
-上下文 1	68.32	68.21	67.66
-上下文 2	69.59	69.32	69.35
-前后缀	66.69	68.29	66.42
-词形 1(istitle/upper/digit)	**70.04**	**69.89**	**69.87**
-词形 2(包含符号)	69.42	**69.85**	69.55
-正则 1	**70.07**	**69.86**	**69.87**
-正则 2	**70.01**	**69.90**	**69.86**
-词性	69.41	**69.79**	69.51
-词形 1 正则 1	69.41	69.67	69.45
-词形 1 正则 2	69.57	**69.83**	69.61
-正则 1 正则 2	69.69	69.63	69.56
-词形 1 正则 1 正则 2	69.51	**70.01**	69.66

通过观察在所有特征的模型上每次减去一个特征的测试结果，可以发现将词形 1（istitle，首字母大写；isupper，所有字母大写；isdigit，数字组成）、正则 1 () 以及正则 2 减去后，模型的召回率、准确率以及 F1 值会有细微的提升。同时在后面的实验中，将这三个特征全都去掉后或者两两去掉，模型性能均有所下降。这是由于特征组合存在较大的不稳定性，同时细微的差别可能是训练模型的偶然误差所导致。通过增加迭代次数，考察这三个特征的稳定性，选择从模型中减去词形 1 特征，作为最终的 CRF 模型。

对最优模型进行分析，获得如表7所示 top positive 和 negative 的影响因子，可以发现前后缀的影响最大，并且既有消极也有积极的影响，这是符合常识的。同时由于其积极影响很大，因此即使其消极影响也较大，但仍然不能贸然移除该特征。

表 7: Top postive 和 Top negative 影响因子

Top positive			Top negative		
6.255084	B-cell_type	word[-2:]=c2	-4.082529	O	word.lower=superantigens
6.182242	B-protein	word[:3]=NLS	-3.797779	O	word[-3:]=NAs
5.939577	O	EOS	-3.561560	I-protein	-1:word[-2:]=-Y
5.427232	B-DNA	word=phoP	-3.501743	I-DNA	-1:word[:2]=p2
5.357717	B-cell_line	word.lower=heterokaryons	-3.036553	O	word[-2:]=Rs
5.058233	B-DNA	word[:3]=5'H	-3.065397	O	word[:2]=GP
5.055145	B-DNA	word[:3]=14q	-3.158875	O	-1:word[-2:]=s-
4.921409	B-DNA	word=fra-1	-3.222680	O	word[-2:]=ag
4.894753	B-protein	word.lower=microtubules	-3.255591	O	word[:3]=onc
4.843137	B-cell_type	word[-2:]=DC	-3.323078	B-protein	word.lower=cyclosporin

(3) 通过在 (2) 中的实验，得到最终的最优特征组合。下面将针对最优特征组合模型进行参数的调优，使用的 python-crfsuite 的模型参数主要包括如表8

表 8: python-crfsuite 的 CRF 训练模型主要参数

参数	说明
feature.possible_states	是否设置发射概率
feature.possible_transitions	是否设置转移概率
c1	超参数
c2	超参数
max_iterations	最大迭代次数
algorithm	训练算法选择

CRF 模型中存在两个超参数 c1 和 c2，因此可以使用网格搜索法进行调参，调参结果如表9

表 9: 使用网格搜索法对 c1 和 c2 进行调参

c1	c2	training loss	R(%)	P(%)	F1(%)
1	1	47397.7	69.31	**70.20**	69.67
	1e-3	40789.7	**70.04**	69.89	**69.87**
0.1	0.1	21061.9	68.0	68.44	68.11
1e-3		14949.8	66.88	67.76	67.2

分析表9，可以发现，当固定 c1 时，c2 越小模型性能越好。而固定 c2，c1 越小，则在训练集上 loss 越小，但在测试集上表现较差，表现为过拟合。因此通过搜索，可以确定 c1 的取值不能太小，而 c2 的取值应当取的略小。最终确定 c1=1，c2=1e-3。之后进行了转移概率和发射概率的研究，发现并不能提高模型性能。

综上所述，CRF 最优模型，应当为：包含除了词形 1 之外的其他特征；模型超参数 c1=1，c2=1e-3。

4.4.2 CRF+Bi-LSTM 模型

由于 CRF+Bi-LSTM 神经网络模型比较复杂，参数比较多，而且每次实验运行时间长。因此，实验过程中通过查阅各种技术博客 [21]，最终从以下几个方面进行模型的调整和分析：

(1) 不同 batch size 的对模型也有很大影响，实验中，根据经验将 batch size 阈值设置 64，然后分别测试 batch size 为 128、64 和 32 的模型性能，得到如表10。当 batch size 过大或者过小时，模型性能都会降低，而当 batch size 为 64 时能够取得最好的性能。

表 10: 字符级别字向量维度对模型的影响

	R(%)	P(%)	F1(%)
batch size = 32	72.56	66.11	69.18
batch size = 64	**75.91**	**67.53**	**71.47**
batch size = 128	61.34	57.31	59.25

(2) 通过实验，可以得到如图11所示不同字符级别字向量维度对模型的影响，可以发现，字符级别字向量的字嵌入维度越高，模型性能越好。当字嵌入维度越高时，越能准确的表达每个单词的内部组成情况（大小写、数字、符号的等），但是训练时间也会与之增长。

表 11: 字符级别字向量维度对模型的影响

	output length=20	output length=64
1978-1989 set	64.73/65.18/64.96	**69.77**/**68.75**/**69.26**
1990-1999 set	76.14/67.58/71.61	**77.47**/**67.96**/**72.41**
2000-2001 set	72.89/66.39/69.49	**77.01**/**68.01**/**72.23**
S/1998-2001 set	73.14/65.5./69.11	**76.79**/**66.77**/**71.43**
Total	72.56/66.11/69.18	**75.26**/**67.87**/**71.33**

(3) 确定模型词向量 (64)、字符级向量 (20-128-64)、BiLSTM(64)、Dense(32)，在此基础上，得到如表12的不同 epoch 下模型的性能表现。可以发现，随着训练时间增长，模型的召回率和 F1 在升高，但以牺牲准确率为代价，因此模型训练时间不能太长，可能会过拟合。

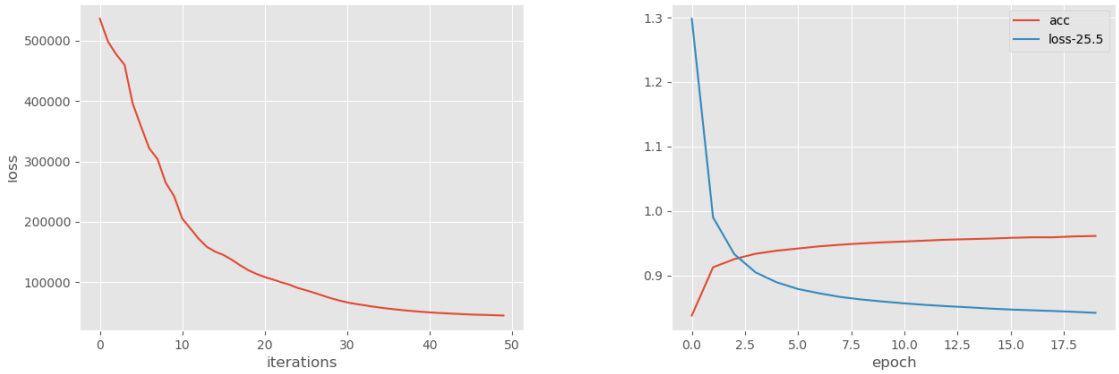
表 12: 不同模型

	epoch=10	epoch=20
1978-1989 set	68.37/67.38/67.87	**69.77**/**68.75**/**69.26**
1990-1999 set	**77.67**/**68.82**/**72.98**	77.47/67.96/72.41
2000-2001 set	76.49/**68.60**/**72.33**	**77.01**/68.01/72.23
S/1998-2001 set	75.74/**66.98**/71.09	**76.79**/66.77/**71.43**
Total	74.57/**67.95**/71.07	**75.26**/67.87/**71.33**

4.4.3 模型对比

在前面的实验分析中，分别对 CRF 模型和 CRF+Bi-LSTM 模型进行了特征选择、模型调优以及性能方面的分析。下面将两个模型进行对比分析，主要分析两者的收敛速度，在训练集上的表现以及在不同时期测试集上的效果。

(1) 对比 LSTM 和 CRF 的收敛速度，如图7。其中 CRF+Bi-LSTM 模型每个 epoch 平均耗时 10mins，CRF 每个 epoch 平均耗时 1s，可以看出 CRF 收敛很快，而神经网络模型收敛速度较慢，这很符合理论常识。因为神经网络通常需要将输入数据映射到高维空间进行系列的矩阵操作，运行速度较慢。



(a) CRF 模型 loss 变化图　　(b) CRF+Bi-LSTM 模型 loss 变化图

图 7: loss 变化图

(2) 选择 CRF 最优模型与 CRF+Bi-LSTM 最优模型进行对比，如表13，给出 CRF 模型和 Bi-LSTM+CRF 模型在测试集上的表现。可以发现 CRF+Bi-LSTM 模型性能明显优于 CRF 模型，其中 CRF+Bi-LSTM 模型具有高召回率、低准确率的特点，而 CRF 模型的召回率、准确率比较稳定。分析原因，主要是由于神经网络具有很强的学习能力，能够充分学习命名实体的特征，因此在其所找到的命名实体基本都是正确的，具有很高的召回率；同时也是由于充分学习特征，而命名实体构成比较复杂多变，因此真正找到符合特征的命名实体少，所以具有较低的准确率。

表 13: LSTM+CRF 和 CRF 在不同时期测试集上的效果

P/R/F1(%)	CRF	CRF+Bi-LSTM
1978-1989 set	63.80/**71.88**/67.60	**69.77**/68.75/**69.26**
1990-1999 set	73.96/**69.45**/71.64	**77.47**/67.96/**72.41**
2000-2001 set	71.38/**69.57**/70.46	**77.01**/68.01/**72.23**
S/1998-2001 set	70.90/**68.16**/69.50	**76.79**/66.77/**71.43**
Total	70.61/**69.16**/69.88	**75.91**/67.53/**71.47**

如表14给出了 CRF+Bi-LSTM 模型和 CRF 模型在不同命名实体上的识别结果,其中 CRF+Bi-LSTM 模型在所有命名实体上识别出来的个数都多于 CRF 模型，并且召回率和 F1 值都高于 CRF 模型，但是仍然是以牺牲准确率为代价，而 CFR 的准确率比较高。

表 14: CRF+BiLSTM 和 CRF 对于不同命名实体的识别结果

num(P/R/F1)(%)	CRF	CRF+Bi-LSTM
proterin(5067)	3801 (75.01/**67.78**/71.21)	**4086** (**80.64**/65.60/**72.34**)
DNA(1056)	683 (64.68/68.16/66.38)	**730** (**69.13**/**71.29**/**70.19**)
RNA(118)	82 (69.49/**66.13**/67.77)	**90** (**76.27**/63.38/**69.23**)
cell_type(1921)	1256 (65.38/**80.77**/72.27)	**1347** (**70.12**/76.58/**73.21**)
cell_line(500)	294 (58.80/53.07/55.79)	**322** (**64.40**/**55.23**/**59.46**)
All(8662)	6116 (70.61/**69.16**/69.88)	**6575** (**75.91**/67.53/71.47)

5 结束语

在本次实验中，先后使用了 CRF 模型和 LSTM+CRF 模型进行了命名实体识别。在使用 CRF 模型时，通过查阅各种资料 [22]，将 CRF 与之前学过的隐马尔可夫模型（HMM）以及最大熵模型（ME）进行了对比，加强了对于此类概率图模型的理解。但是没能自己实现 CRF，所有对于 CRF 的学习算法了解甚少，在以后的学习中，可以加强对这方面的学习，争取能够实现相关的算法模型。在 CRF 模型的特征选择过程中，使用到了比较多的特征，学习了许多对比分析、控制变量的科学方法，对于研究思路以及今后的实验有着指导性的意义。

关于神经网络,虽然之前选修过深度学习的相关课程,但是在本次实验中使用 LSTM+CRF 模型时，仍然遇到一些困难。实验过程中通过学习各种技术博客以及查阅相关书籍，最终在耗费了 2 天的时间后，成功搭建起了神经网络。但是在运行过程却遇到较大阻力，主要在于运行时间过于长。在笔记本上尝试使用 GPU 运行，但效果比较差，最终选择租用较高配置的云服务器运行。之后又用了大量的时间来调参，但效果并不是特别的理想。同时模型也存在很多可以改进的地方，比如从增加向量维度、使用 CNN 等方面尝试训练更好的词向量和字符级别的字向量 [23]，应该能获得更好的结果，但是由于时间不足，没能进行相关的尝试，以后可以在这方面进行相关的研究学习。然后就是对于神经网络的实现细节不甚了解，搭建的神经网络模型不够健壮，层与层之间的连接并不太好，在以后的学习中可以进一步完善。

通过实验，也对加深了对课程相关知识的掌握。实验中用到词性标注的技术，就更深刻的了解到了词性标注技术和实体识别技术之间的联系。两种技术本质上是相同的，都可以转化为序列标注问题，只是词性标注更加细致，而实体识别只需要对其中的实体进行标注即可。因此类似 CRF、HMM 等概率图模型以及 RNN 等神经网络这些方法，两者都是通用的。同时，对于之前学习过的分词技术，也有着同样的处理思路，可以通过标注每个词在分词中的位置转化为序列标注问题。

除了上述关于模型方法等方面的收获外，此次实验最大的收获便是论文的撰写。由于指导书要求以科技论文的组织方式书写报告，为了达到较好的效果，阅读了许多论文，通过比

较分析最终确定了本文的组织方式。而整个实验过程也基本按照这个思路进行，实现了一个完整的实验研究、论文撰写的流程。关于论文的撰写，则第一次尝试使用 LaTex，体验还是很好的，相信对为以后的学习研究以及毕业论文撰写会有着不小的帮助。

总的来说，整个实验过程，学习到很多知识，对之后的学习奠定了一定的基础；也认识到许多不足，这也为未来的学习研究指明了方向。

参考文献

[1] Ralph Grishman and Beth Sundheim. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, volume 1, 1996.

[2] Christine Thielen. An approach to proper name tagging for german. *arXiv preprint cmp-lg/9506024*, 1995.

[3] K Jin-Dong. Report on bio-entity recognition task at bionlp/nlpba 2004, 2016.

[4] 王宁, 葛瑞芳, 苑春法, 黄锦辉, and 李文捷. 中文金融新闻中公司名的识别. *中文信息学报*, 16(2):1–6, 2002.

[5] Zhao Jian. Research on conditional probabilistic model and its application in chinese named entity recognition. *Harbin Institute of Technology*, 2006.

[6] 陈霄, 刘慧, and 陈玉泉. 基于支持向量机方法的中文组织机构名的识别倡. *计算机应用研究*, 25(2), 2008.

[7] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[8] Burr Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 104–107. Association for Computational Linguistics, 2004.

[9] 姜文志, 顾佼佼, and 丛林虎. Crf 与规则相结合的军事命名实体识别研究. *指挥控制与仿真*, 33(4):13–15, 2011.

[10] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[11] Liyuan Liu, Jingbo Shang, Frank Xu, Xiang Ren, Huan Gui, Jian Peng, and Jiawei Han. Empower sequence labeling with task-aware neural language model. *arXiv preprint arXiv:1709.04109*, 2017.

[12] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

[13] 韩雪冬 and 周彩根. 条件随机场理论综述. 中国科技论文在线, 2010.

[14] Joel Nothman, Tara Murphy, and James R Curran. Analysing wikipedia and gold-standard corpora for ner training. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 612–620. Association for Computational Linguistics, 2009.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[16] 陈彦妤 and 杜明. 基于 crf 和 bi-lstm 的保险名称实体识别. 收藏, 3:026, 2018.

[17] https://python-crfsuite.readthedocs.io/en/latest/.

[18] https://github.com/keras-team/keras-contrib.

[19] https://www.nltk.org/.

[20] http://www.nactem.ac.uk/tsujii/GENIA/ERtask/report.html.

[21] https://www.depends-on-the-definition.com/sequence-tagging-lstm-crf/.

[22] 刘建伟, 黎海恩, and 罗雄麟. 概率图模型表示理论. 计算机科学, 41(9):1–17, 2014.

[23] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.