哈 尔 滨 工 业 大 学

国 家 示 范 性 软 件 学 院

本科毕业设计(论文)中期报告

题 目：基于深度学习的命名实体识别系统设计与实现

专　　业	软件工程
学 生 姓 名	乐远
学　　号	1143710316
联 系 方 式	13069875982
年　　级	2014 级
实 习 基 地	深圳证券交易所
基地指导教师	许保勋
联 系 方 式	13069875982
校内指导教师	郭 勇
联 系 方 式	13030000672
中 检 日 期	2018.03.31

哈尔滨工业大学软件学院

目　录

1. 毕业设计（论文）内容概述

1.1 项目来源及开发目的和意义

1.1.1 项目来源

本项目主要来源于我在在深圳证券交易所实习阶段所参与的《证券金融知识图谱》项目以及许保勋许博士的指导。

1.1.2 项目开发目的和意义

互联网+时代的到来标志着互联网从一个工具变成了一个基础性的设施，在互联网+时代，万物通过互联网进行互联，互联网的基础性地位日显重要，已经渗透到包括金融、物流、电子商务、工业生产等各个领域。互联网以信息作为作为其载体及表现形式的特性，与金融行业有天然的融合性。金融行业从本质上而言，就是用不同的数字与信息去表达金融资源的时间与空间特性，通过对信息进行处理，完成不同金融资源的时间及空间的匹配，以达到资源效用最大化的目的。

金融证券行业对信息的分析与处理方法的探索从来没有停止过。以股票市场为例，早期受到分析手段及资讯传导速度的限制，人们以分析结构化数据，例如股票的成交量、成交价格为主；在公司的基本方面，则以分析公司的财务结构数据为主。在报业时代，受信息更新速度、传播速度的影响，通过对非结构化的文本数据包括并不多；与此同时，报业时代产生的数据量并不大，由人工分析足以满足业务应用需求。在信息时代，一方面随着互联网时代的到来，资讯的生产方由专业媒体变成了大众，各类关于公司、市场的信息由不同的人士生成并发布，数据量空前丰富；另一方面，空前丰富的数据体量使得人工分析变得越来越不现实，信息技术的成熟、应用成本的降低使得将信息技术应用于金融非结构化数据的分析服务成为可能。在这个时期，证券行业通过搭建各类分析平台对结构与非结构化数据进行采集与分析。

然而上述对信息的分析方法仍然存在缺陷。首先，目前互联网已经进入到

了互联网+时代，万物互联已经成为主流，而上述的信息分析方法将一个个信息点进行孤立的分析，形成一个个信息分析孤岛，其表现形式为对单一问题、单一信息分析较为全面，但对多个问题、多个信息的关联分析等能力较为欠缺，分析结果零散，查询结果不够智能，只能就查询者的某个问题回答相应的答案，而不能够就问题所描述的知识结构完整全面的战士给查询方。这些等等问题催生了 Google 公司推出的知识图谱在金融证券领域的应用。

自 2012 年谷歌将知识图谱成功应用到搜索引擎以来，知识图谱在学术界和工业界收到了广泛关注。知识图谱的本质是由概念、实体以及实体之间的关系构成的语义网络。知识图谱的构建主要是将零散的结构化、半结构化和无结构化数据通过信息抽取、信息融合等技术处理成集中的结构化数据，并通过图的方式表达实体与实体之间的复杂关系，方便上层应用系统从整个知识系统的角度去分析复杂的逻辑推力问题。构建金融证券领域的知识图谱需要从基于互联网平台的股吧、论坛、门户网站、微信、微博、公告、研报、招股相关文档等等结构化或非结构化的数据中进行信息抽取、信息融合，达到人、公司、产品、行业的“万物互联”（如图 1-1），从而提高行业信息利用的精准度和可信度，以及广度。通过证券行业知识图谱将所有重点相关联的行业、版块、公司、股票以及个人进行影响价值，对上述信息可能产生的正面或者负面影响进行实时的分析并得出相应的结论，使得机构可以先于市场其他参与者发掘出潜在关联方并全面的分析出事件波及影响层面，从而快速作出投资决策实现盈利或止损。因此研究金融领域证券知识图谱的构建具有重大意义。

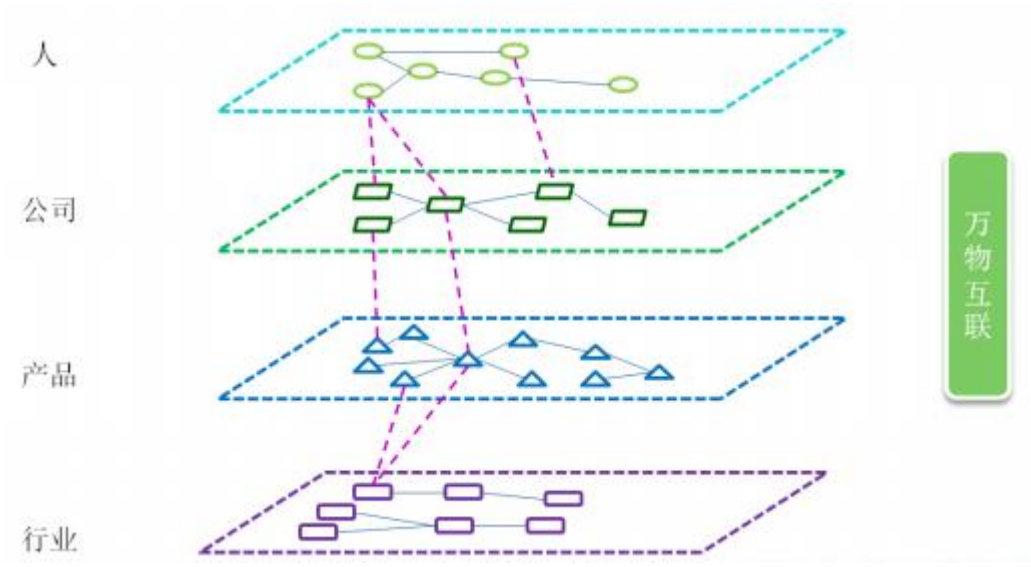

图 1-1 证券知识图谱万物互联图

如果把金融领域证券知识图谱构建分为知识构建、知识计算、知识存储、知识应用四大部分（如图 1-2），那么知识构建应该是最核心基础的一大部分，即怎么从海量文本中得到行业图谱。

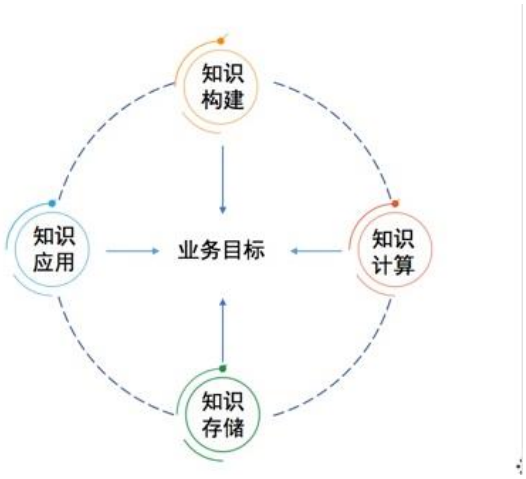

图 1-2 金融知识图谱构建组成部分

而金融领域证券知识图谱的知识构建最核心的两大技术就是命名实体识别（Named Entity Recognition）和关系抽取（Relation Extraction），而命名实体识别就是如何从文本数据中抽取概念、实体、关系和属性并进行消歧、对

齐、融合。

金融行业面对的数据来源多样、结构复杂，其中既包括来自互联网舆情、监督机构的合规要求、内部报告等文本数据，财务、行研等版结构化数据，以及上百个业务系统产生的海量结构化数据，在抽取实体、关系、属性时，会面临消歧、对齐、融合等难点。因此设计一套适合金融领域知识图谱构建的命名实体识别算法，使得构建的知识图谱更精确、效率更高，具有重大的使用价值。

1.2 主要开发任务

系统主要是完成命名实体识别的任务， 如图 1-3，本命名实体识别系统主要是从招股书、年报、公司报告、新闻等半结构化和非结构化文本数据中批量自动识别实体名（包括人名、地名、机构名、专有名词等等）、时间表达式（包括时间、日期等等）、数量表达式，输出 BIO 标注结果集，并且需要进行消歧、对齐、融合等等。

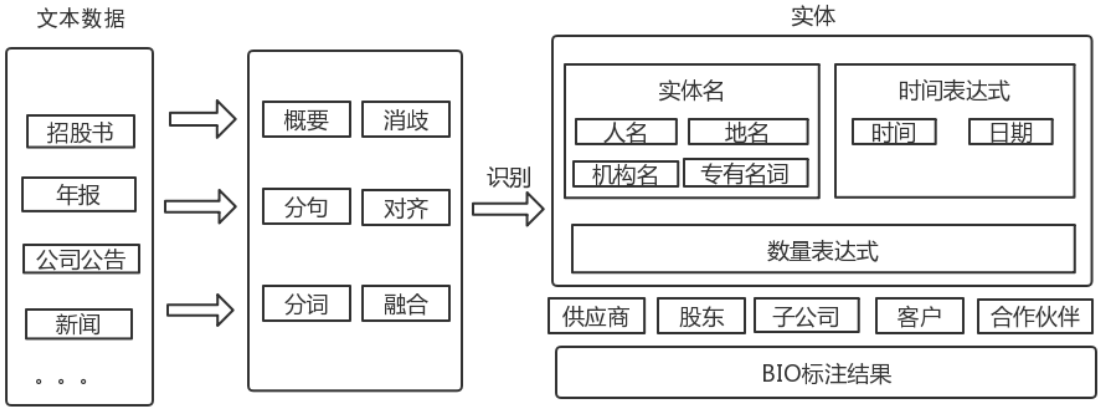

图 1-3 命名实体识别系统主要功能示意图

系统分为三大模块， 主要包含预处理、命名实体识别、后处理三大部分，预处理又包含正文提取与概要、分句、分词模块，命名实体识别包含词向量、双向 LSTM、CRF 模块，后处理又包含消歧、对齐、融合模块。具体的

系统功能结构图如图 1-4 所示。

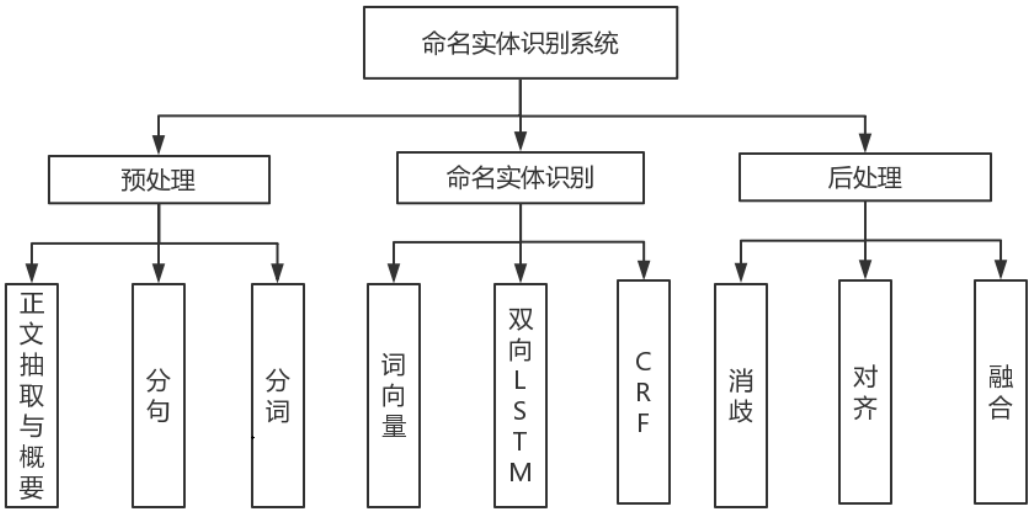

图 1-4 命名实体识别系统功能结构图

1.3 本人所承担任务（模块）说明

本人主要是实现命名实体识别的核心算法模块和后处理模块!

（1）命名实体识别包括分句、分词以及词向量、双向 LSTM、CRF 等

- 需要对输入的句子使用词向量表示，使用 word2vec 训练中文词向量
- 将句子输入到双向 LSTM 得到上下文信息
- 使用条件随机场 CRF 对句子进行标注，使用 BIO 标注集

（2）后处理模块包括消歧和对齐。

- 消除部分实体的歧义
- 融合部分实体

1.4 开发环境和开发工具

1.4.1 开发语言

本系统主要使用 python 语言开发，特别是使用了 tensorflow 深度学习框架。

1.4.2 开发工具

本系统使用 PyCharm2017 作为主要开发平台，调试时使用 PyCharm2017 集成的调试工具。

如表 1-1 所示，本系统的开发使用到了这些开发工具：

表 1-1 开发工具表

工具类别	工具名称	作用
集成开发环境	PyCharm2017	程序最主要的开发、调试平台
英伟达显卡	GTX 750	加速深度学习的计算
版本控制软件	Github	代码备份，版本控制

1.4.3 开发环境

（1）系统开发环境

操作系统：ubuntu16.04

处理器： Intel(R) Xeon(R) CPU E5-2660 0 @ 2.20GHz

安装内存：64GB

系统类型：64 位操作系统

（2）系统运行环境

操作系统：Windows Server 2008 R2 Standard

处理器： Intel(R) Xeon(R) CPU E5-2660 0 @ 2.20GHz

安装内存：288GB

系统类型：64 位操作系统

1.5 项目原定进度安排

项目进度及毕业设计（论文）工作安排见表 1-2。

表 1-2 项目进度及毕业设计（论文）工作计划表

起始时间	完成时间	计划工作内容	备注
2017.07.17	2017.09.01	跟随导师了解项目，毕设选题	已完成
2017.09.02	2017.09.30	完成项目需求与可行性分析	已完成
2017.10.01	2017.10.12	完成概要设计	已完成
2017.10.13	2017.10.31	阅读相关论文，调研领域算法	已完成
2017.11.01	2017.11.10	撰写开题报告，准备开题答辩	已完成
2017.11.11	2017.11.20	完成开题报告，答辩 PPT	已完成
2017.11.21	2017.11.30	收集相关数据，标注数据集	已完成
2017.12.01	2018.01.31	实现深度学习算法，训练，调参	已完成
2018.02.01	2018.03.10	初步完成 NER 的基本识别功能	已完成
2018.03.11	2018.03.20	准备中期答辩	在进行
2018.03.21	2018.04.21	继续调整优化算法，调整模型结构	未完成
2018.04.22	2018.05.22	完善算法的设计细节	未完成
2018.05.23	2018.06.15	用多组大规模数据进行训练，测试	未完成
2018.06.16	2018.06.20	完善最后的一些细节	未完成
2018.06.21	2018.07.01	撰写、修改论文，参加毕业答辩	未完成

2. 中期完成情况说明

2.1 预定计划的执行情况

设计工作基本上按照进度安排进行，其中有小部分改动是命名实体识别只识别人名、地名、组织名（包括公司名），不再识别时间、日期等数据，因为这个对关系抽取工作以及证券知识图谱的构建没有什么太大的作用，而且实现很难！这是和小组一起讨论决定的。

2.2 中期工作说明及成果汇报

2.2.1 系统详细设计

（1）系统流程图

命名实体识别系统的各个模块功能之间是紧密联系的，前者的输出即是后者的输入，系统先读入文本、分句，分词并判断是否借助字典，然后进行命名实体识别得到 BIO 标注结果，最后判断是否需要进行后处理，该命名实体识别系统程序流程图如图 2-1 所示。

文本数据的读入需要我们从结构化数据和非结构化数据中提取出文本的主要信息，然后分句、分词，分词借助字典，我们使用开源的结巴分词等工具，然后运行实体识别算法得到 BIO 标注结果，这也是最为核心的一部分。最后对 BIO 的标注统计结果判断是否进行消歧、对齐、融合，避免信息的冗余。

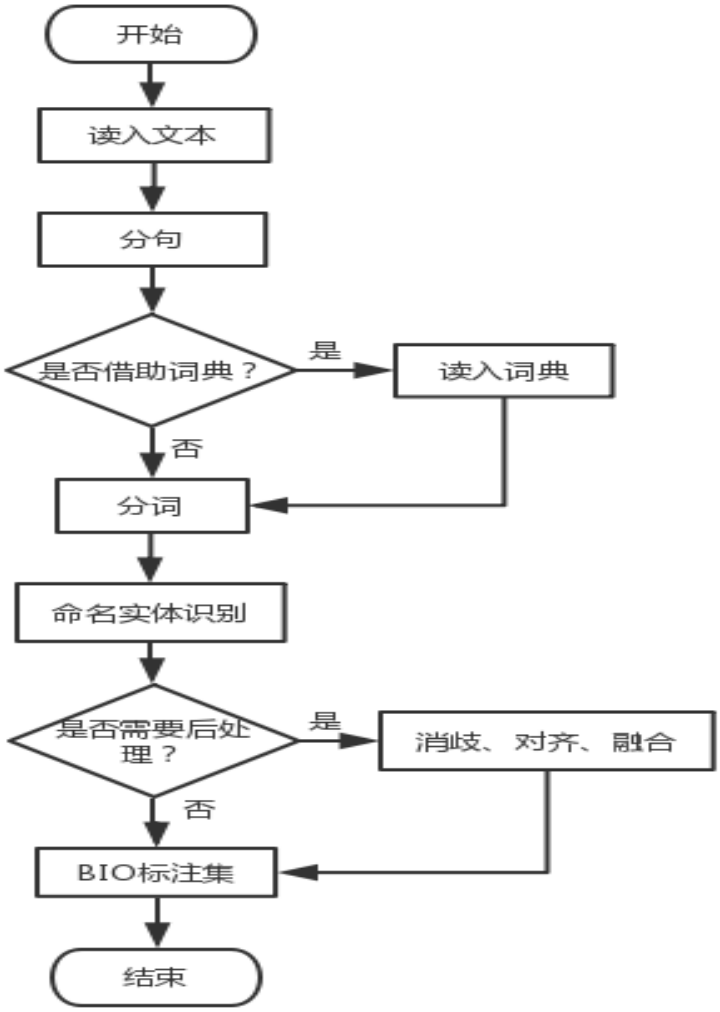

图 2-1 命名实体识别系统流程图

（2）系统架构设计

如图 2-2，命名实体识别系统可以划分为以下几个部分来实现。用户界面负责所有的人机交互，用户通过用户界面上的菜单、按钮等控件来调用系统的 4 大核心模块：文本预处理模块、实体识别模块、文本后处理模块、日志模块。

文本预处理模块可以读入文本数据、分句、分词，是实体识别系统的首要模块。实体识别模块可以训练、调用命名实体识别算法模型，对给定的文本进

行实体识别给出 BIO 标注结果，这是实体识别系统的核心模块。BIO 标注集和相关的模型参数都以文本的形式存储在硬盘上。日志模块记录用户使用该命名实体识别系统的工作日志，日志数据也是以文本形式存储在硬盘上。

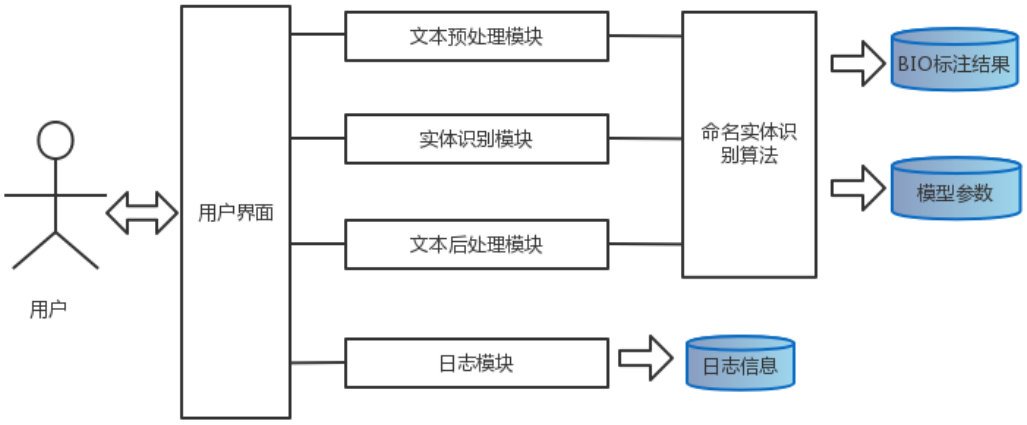

图 2-2 命名实体识别系统架构图

2.2.2 系统算法详细设计

我们采用了基于深度学习的 BiLSTM+CRF 的命名实体识别方法来完成到金融领域的中文命名实体识别任务，并更改了部分神经网络结构，以适应金融领域。 下面详细介绍我们的算法各个网络结构层。

（1）总体架构

图 2-3 展示了基于 LSTM 单元的完整的命名实体识别的深度学习模型整体框架，主要分为 4 个步骤：

1）经过分词后的完整的句子序列首先进入词向量层，词向量层维护了一个参数矩阵，称为词向量查找表，输入的句子能够通过这个矩阵转换为对应词的词向量的序列。我们使用 word2vec 来训练词向量。

2）词向量的序列根据设定的参数窗口大小将词向量进行连接，设窗口大

小为 k，序列长度为 N，则得到 N-k+1 的连接序列，作为 BiLSTM 层的输入序列。

3）利用随机初始化对 BLSTM 层的多个参数矩阵进行初始化，步骤 2 中得到的输入序列进入 BLSTM 层，即同时输入到正向 LSTM 层与反向 LSTM 层进行模型计算和训练。为了防止出现过拟合的情况，可以在 LSTM 层的输入和输出部分加入 dropout 机制，最后拼接得到的两个方向的 LSTM 输出得到整个 BLSTM 层的输出序列作为隐含层的输入。

4）隐含层的序列输出经过与参数矩阵相乘，得到转移概率的参数矩阵，维度为序列长度*输入标记的种类个数，用来进行最终正确路径的搜索。模型训练时利用极大似然估计法进行概率计算，利用维特比算法进行测试时序列的解码。

（2）长期短期记忆网络 LSTM

长期短期记忆网络(LSTM)的提出是为了解决 RNN 的长距离依赖问题，它加入了内存单元(Memory Cell)机制和门(gate)机制，控制输入到内存单元的比例以及前一个状态到遗忘门(forget gate)的比例，数学公式表示如下：

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (2\text{-}1)$$

$$c_t = (1-i_t)\odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (2\text{-}2)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (2\text{-}3)$$

$$h_t = o_t \odot \tanh(c_t) \quad (2\text{-}4)$$

其中 X=（$x_1,x_2,\ldots,x_n$）表示一个句子，h=（$h_1,h_2,\ldots,h_n$）表示 LSTM 对于每个输入所对应的某种信息表示。

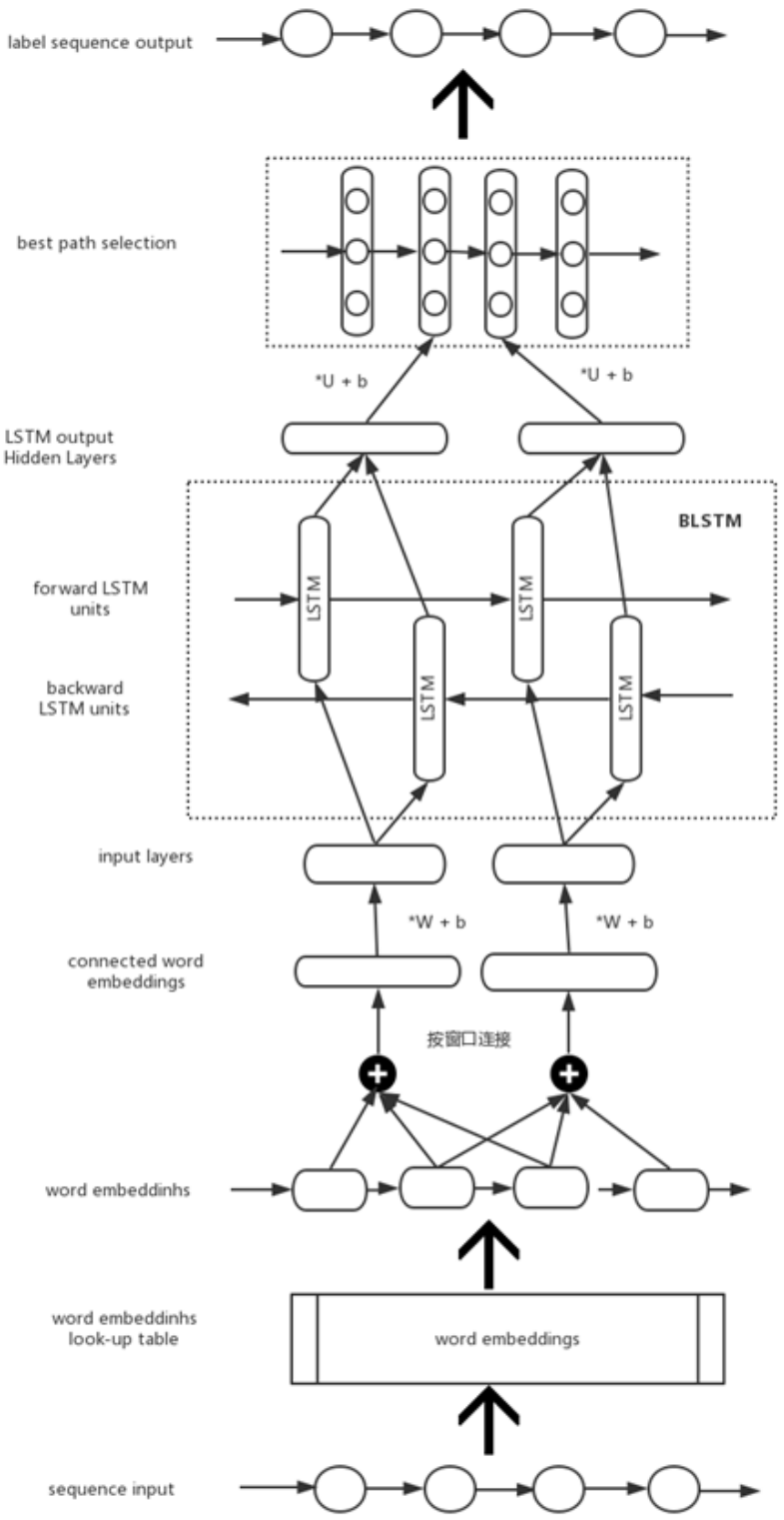

图 2-3 基于 LSTM 的 NER 模型整体框架

(3) 窗口连接

如果使用单一方向的 LSTM 单元，整体模型在上下文信息处理上，事实上是缺乏后文信息的。而双向的 LSTM(Bidirectional LSTM, BTSTM)同时联结了上文和下文两个方向的额 LSTM 单元，能够捕捉到上文和下文信息，但是整个模型的计算量是单一方向的 LSTM 的双倍。而在 NER 等类似的词法分析任务中，有时候后文的信息会更加重要，因此使用了类似于 CRF 等模型的窗口方法，使得模型在不大幅度增加计算量的情况下，可以获取指定窗口大小的前后文信息。在图 2-3 中“按窗口连接”可以看到。

（4）Dropout 机制

在训练神经网络的时候，由于网络参数过多，很容易出现过拟合的情况。Dropout 机制可以有效的防止过拟合，因此我们在 LSTM 单元的输入层和 LSTM 单元的隐含层输出两端加入了 Dropout 机制。

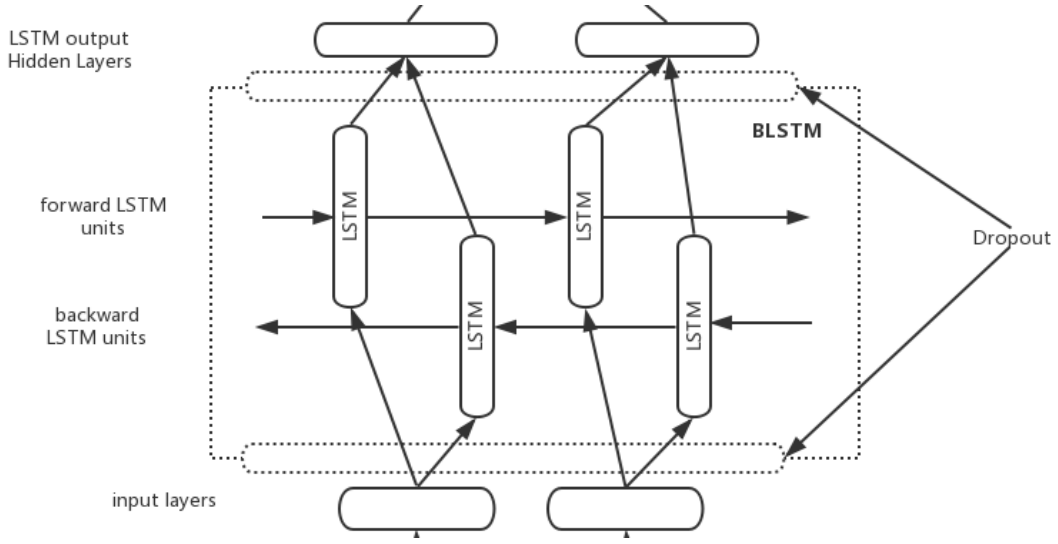

图 2-4 NER Dropout 机制加入位置

（5）引入转移代价的代价计算

借鉴于 Collobert 等提出的模仿 CRF 等模型，采用计算路径的转移概率从

而计算出正确路径的方法，这里我们也采用这种方法。

转移概率的计算位于 LSTM 单元将隐含层得到序列输出之后，输出矩阵乘以维度大小为（隐含层大小*标注符号集大小）的参数矩阵，得到了转移概率的分数矩阵，其维度大小为（序列长度 n*标注符号集大小 k）记作矩阵 P。这个矩阵的实质是输入序列中不同位置标注为不同待标注标记的分数矩阵。在模型的训练的过程中首先使用极大似然法进行优化，而在测试的过程中，会采用维特比算法进行最优路径的解码和选择。

对于每个输入句子 X=（$x_1,x_2,\ldots,x_n$），通过该模型得到一个预测序列 $y=(y_1,y_2,\ldots,y_n)$,定义分数函数为：

$$s(X,y)=\sum_{i=0}^{n}A_{yi,yi+1}+\sum_{i=1}^{n}P_{i,yi} \tag{2-5}$$

其中矩阵 P 是双向 LSTM 的输出，维度为 n×k，$P_{i,j}$ 表示句子中的第 i 个词标注为第 j 个标签的分数；矩阵 A 是代价转移矩阵，$A_{i,j}$ 表示从标签 i 转移到 j 的转移分数，维度是(k+2)×(k+2)，y_0 和 y_{n+1} 表示一个句子的开始和结束标签，所以是 k+2。

通过 softmax 得到将句子 X 标注为 y 的概率如下：

$$p(y|X)=\frac{e^{s(X,y)}}{\sum_{\tilde{y}\in Yx}e^{s(X,\tilde{y})}} \tag{2-6}$$

其中 Yx 表示对于句子 X 所有可能的序列标注。使用极大似然估计得到：

$$\log(p(y|X))=s(X,y)-\log(\sum_{\tilde{y}\in Yx}e^{s(X,\tilde{y})}) \tag{2-7}$$

通过训练模型，最大化似然函数，使得模型能够产生更准确的输出序列，得到相应的模型参数。在解码时我们将句子标注为具有最大分数的标注序列，计算如下：

$$y^{*}=\arg\max_{\tilde{y}\in Yx}s(X,\tilde{y}) \tag{2-8}$$

2.2.3 系统实现与结果

（1）数据集

我们使用的数据集是第三届 SIGHAN Bakeof 中文命名实体识别任务的 MSRA 数据。这个数据包含三种类型的实体：人名、地名、组织名。然后我们还结合金融领域的特殊性标注了 900 多个句子的数据，合并起来作为最终的数据集。我们将数据集分为了三部分：训练集、验证集、测试集，下表展示了各个数据集的大小：

表 2-1 数据集大小

训练集	验证集	测试集
6131KB	686KB	1373KB

（2）算法模型执行过程

命名实体识别算法执行过程如下：首先读入文本句子，判断是否使用词向量，如果不使用词向量就随机初始化每个词的词向量，否则就加载相应的词向量来表示句子；在判断是否引入其他特征如分词特征，得到特征向量并拼接到句子向量中，然后再依次输入到 BiLSTM 层和 CRF 层，得到相应的概率表示，最后采用我们的解码算法训练模型或者给句子进行标记。我们按照以下流程来执行我们的算法，如下图：

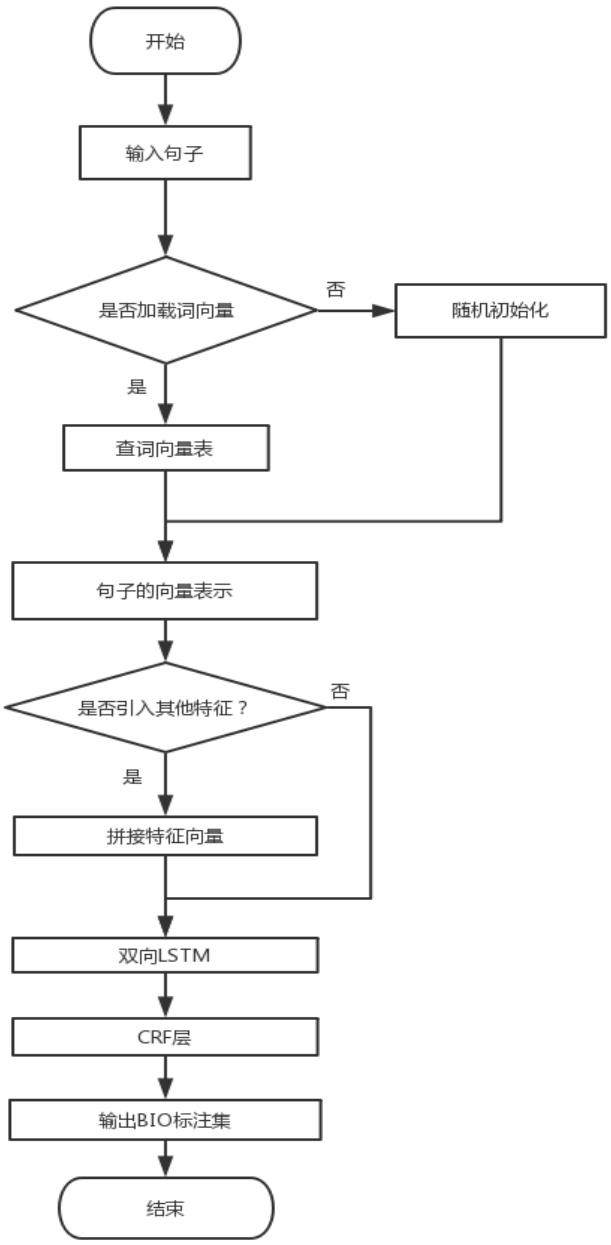

图 2-5 BiLSTM+CRF 算法程序流程图

（3）训练方法

我们使用 python 的 tensorflow 深度学习框架实现了 BiLSTM+CRF 模型， 在 MSRA 训练数据集上训练模型，每训练一个 Epoch，就在验证集上测试模型的准确率、召回率、F 值，然后在测试数据集上测试得到 F 值作为模型

最终的效果。我们使用 tensorflow 的 tensorboard 监控模型的训练，查看模型的 graph 如图 2-6 所示，这个就是我们所实现的模型结构。

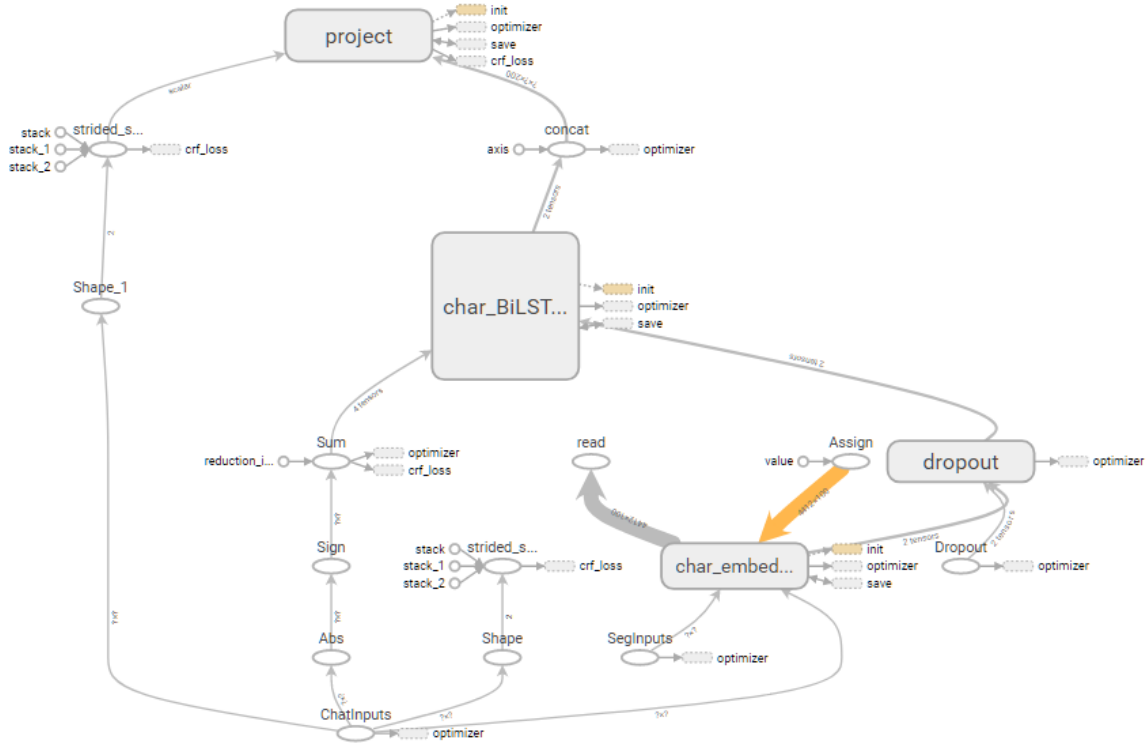

图 2-6 BiLSTM+CRF 的模型图

我们训练模型使用了很多组不同的参数，依次比较使用词向量和不使用词向量的结果；对比是否使用 dropout 的结果，并探寻 dropout 设置为多少效果最优；然后依次找到 hidden state 的最优维度、最优的 batch_size，最优的学习率，最后找到到最优的参数组合，下面的流程图展示了我们调参的一个过程，通过该方法我们找到了还不错的参数组合。

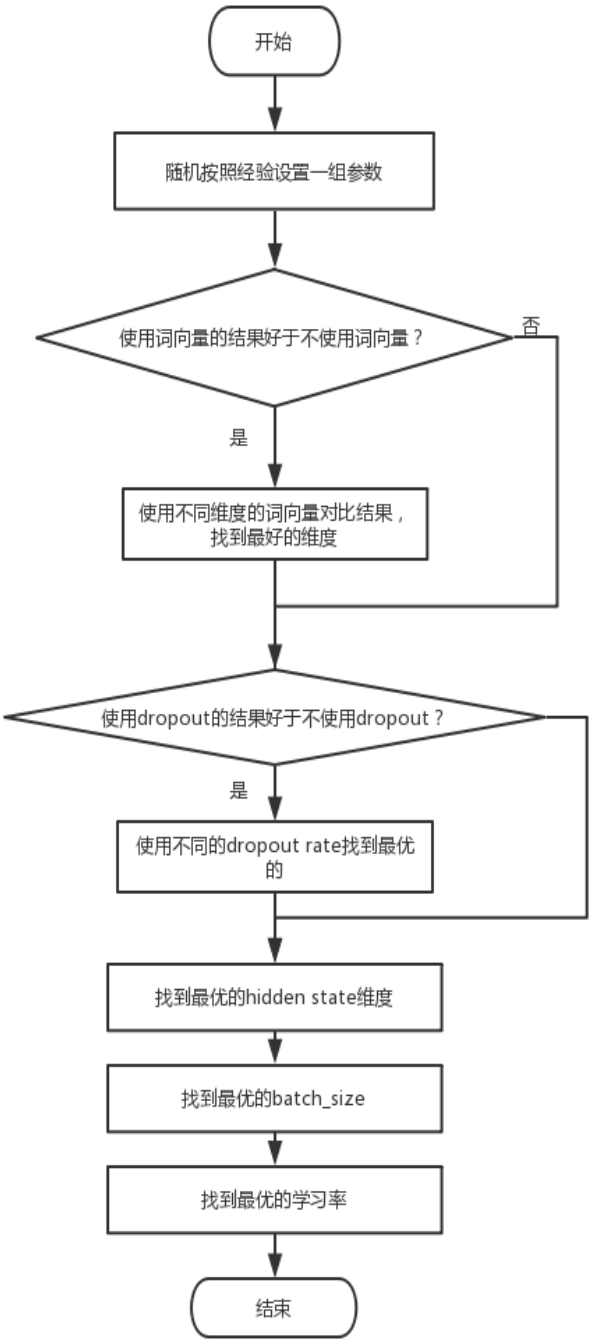

图 2-7 模型参数调优流程图

我们的最优的模型设置字符向量的维度为 100，LSTM 隐藏层维度为 100，batch_size 为 20，dropout rate 使用 0.5，使用 Adam 优化算法，学习率设置为 0.001，训练过程中损失函数值变化如下：

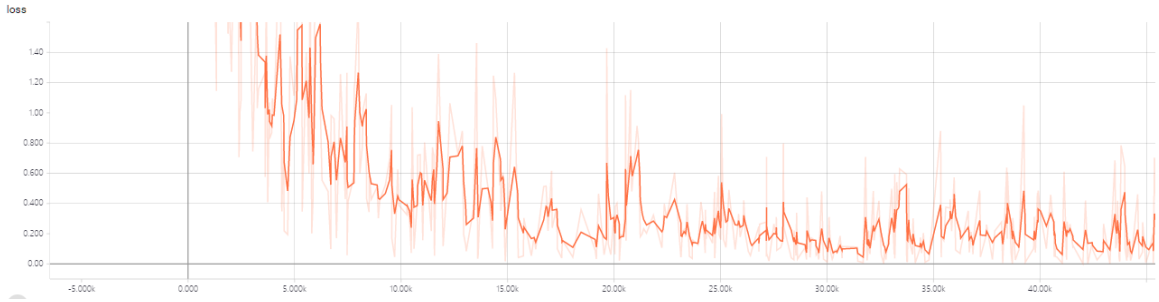

图 2-8 训练过程损失函数 loss 值变化图

模型在验证集和测试集的 F1 值随着训练过程变化如下：

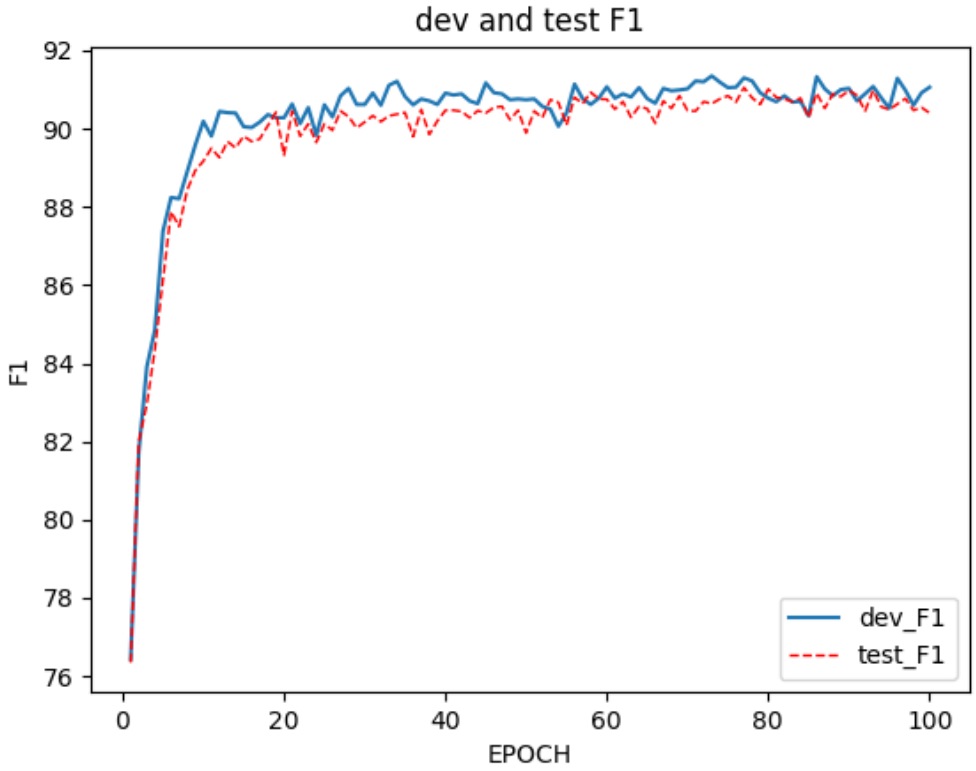

图 2-9 模型训练过程的 F1 值变化图

从图 2-6 和图 2-7 我们看到我们的模型随着训练过程的进行，损失函数值比较平滑的下降，而且模型在验证集和测试集上的 F1 值也上升的比较平滑，表明我们的模型一直在改善，效果越来越好！

（4）词向量

我们使用 pretrained 的方式来训练词向量，使用 word2cvec 在 1.4GB 的 wiki 中文语料库上训练中文字向量，分别训练了 50 维、100 维、200 维的字向量。然后使用不同的词向量维度来训练神经网络以及不使用词向量来训练我们的神经网络，并进测试给出相应的结果。

表 2-2 不同维度词向量训练结果

词向量维度	F1
不使用词向量	88.22
50	89.21
100	90.75
200	90.22

我们发现词向量对于命名实体识别的效果确实有很大的改善，大约能提高 1-2 个百分点。而且我们发现 100 维度的词向量效果是最好的。

（5）dropout 技术

为了防止模型训练过程中过拟合，我们采用了 dropout 技术。得到句子的词向量后，使用 dropout 在输入到双向 LSTM。我们对比了下是否使用 dropout 以及使用 pretrained 的词向量的结果，发现

表 2-3 LSTM+CRF 几种训练的模型对比

训练的几种模型	F1
random	86.24
random + dropout	87.03
pretrain	87.98
pretrain + dropout	90.75

（6）最优的模型结果

1）我们最优的模型在测试集上达到了 90.75 的 F1 值，具体在几种类别的实体上的结果如下：

表 2-4 最优模型的测试结果表

实体类别	P	R	F
所有实体	91.03	90.48	90.75
地名(LOC)	91.28	92.48	91.88
组织名(ORG)	87.65	84.49	86.04
人名(PER)	94.37	95.36	93.97

可以看到我们的模型在人名和地名的识别上 F 值还是相当高的，特别是人名。不过在组织名上的 F 值就没有那么高了，这是由于组织名是几种实体名中最为复杂的，也是最难的一部分。

2）我们从测试数据上抽取了一部分样例，来更直观的显示我们的模型的效果：

第一个句子：

sent: 同一天，约 旦 议 会 发表声明指出，以 色 列 政府批准
true: O O O O B-ORG I-ORG I-ORG I-ORG O O O O O O O B-LOC I-LOC I-LOC O O O O
pred: O O O O B-ORG I-ORG I-ORG I-ORG O O O O O O O B-LOC I-LOC I-LOC O O O O

耶 路 撒 冷 扩建计划旨在扼杀有关各方为中 东 和平所作出的种种努力。
B-LOC I-LOC I-LOC I-LOC O O O O O O O O O O O O B-LOC I-LOC O O O O O O O O O O
B-LOC I-LOC I-LOC I-LOC O O O O O O O O O O O O B-LOC I-LOC O O O O O O O O O O

第二个句子：

sent: 韩 国 财 政 经 济 部 长官李 揆 成 0 0 日说，
true: B-ORG I-ORG I-ORG I-ORG I-ORG I-ORG I-ORG O O B-PER I-PER I-PER O O O O O
pred: B-ORG I-ORG I-ORG I-ORG I-ORG I-ORG I-ORG O O B-PER I-PER I-PER O O O O O

韩 国 经济今年可能出现0 0 年来首次负增长，失业人数最高可能达到0 0 0 万。
B-LOC I-LOC O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O
B-LOC I-LOC O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O

第三个句子：

sent: 霍 华 德 在会见中说，董 建 华 先生访澳 必将促进澳 大
true: B-PER I-PER I-PER O O O O O O B-PER I-PER I-PER O O O B-LOC O O O O B-LOC I-LOC
pred: B-PER I-PER I-PER O O O O O O B-PER I-PER I-PER O O O B-LOC O O O O B-LOC I-LOC

利 亚 和香 港 在贸易、投资方面的合作，增进彼此在人员方面的往来。
I-LOC I-LOC O B-LOC I-LOC O O O O O O O O O O O O O O O O O O O O O O O O O O
I-LOC I-LOC O B-LOC I-LOC O O O O O O O O O O O O O O O O O O O O O O O O O O

第四个句子：

丹 麦 队 的表现似乎不怎么好，在热身赛中输了两场，尤其是以0 ：0 负于
B-ORG I-ORG I-ORG O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O
B-ORG I-ORG I-ORG O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O

未能进入世界杯决赛的瑞 典 ，被看作是" 近来最糟糕的失败" 。
O O O O O O O O O O B-LOC I-LOC O O O O O O O O O O O O O O O O O O O O
O O O O O O O O O O B-LOC I-LOC O O O O O O O O O O O O O O O O O O O O

第五个句子：

现在搞农业除了机械化，还要靠信息化，每天在劳动之后还要在家里的计算机

O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O

O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O

.面前坐好一阵子，浏览信息、寻找客户、了解外面的世界。

O O O O O O O O O O O O O O O O O O O O O O O O O O

O O O O O O O O O O O O O O O O O O O O O O O O O O

第六个句子：

安 馆长介绍，海 大 图 书 馆 的开放式办馆模式特点有三：

B-PER O O O O O B-LOC I-LOC I-LOC I-LOC I-LOC O O O O O O O O O O O O O

O O O O O O B-LOC I-LOC I-LOC I-LOC I-LOC O O O O O O O O O O O O O

第一，所有图书资料向社会读者开放，成年人想看书就进门，不看身份，不验证件。

O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O

O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O

从上面的句子我们可以看到，我们的模型的准确率还不错， 在前面四个句子中几乎预测正确，人名、地名、组织名都准确的识别出来了。第五个句子没有实体，所以我们的模型将其都标注为 O。不过我们的模型也有所欠缺，比如第六个句子没有将“安馆长”中“安”识别出来为人名，因为这个确实很难识别。不过总体来看，整体效果还不错。

（7）与其他算法模型的比较

我们还将我们的模型和其他模型作了比较，其他模型也是使用了第三届 SIGHAN Bakeof 中文命名实体识别任务的 MSRA 数据，我们分别比较了是被人名、地名、组织名以及所有种类试题的准确率(precision)、召回率(recall)、F 值(F score)，如下表所示：

表 2-5 NER 模型对比

模型	PER-F	LOC-F	ORG-F	P	R	F
Zhou2006	90.09	85.45	83.10	88.94	84.20	86.51
Chen2006	82.57	90.53	81.96	91.22	81.71	86.20
Zhou2013	90.69	91.90	86.19	91.86	88.75	90.28
Zhang2006*	96.04	90.34	85.90	92.20	90.18	91.18
our models	93.97	91.88	86.04	91.03	90.48	90.75

第一个模型使用了单词级别的条件随机场 CRF，并且添加了很多的手工特征模板，在 MSRA 数据集上达到了 86.51 的 F1 值。第二个模型使用了字符级别的 CRF 模型，达到了 86.20 的 F 值。第三个模型使用了一个全局的线性

模型来识别和分类中文命名实体，并且混合了 10 个上下文特征模板，达到了 90.28 的 F 值。第四个模型使用了最大熵模型并且结合了很多额外的知识，例如人名列表、组织名字典、地址关键词列表等等，达到了 91.18 的特征值。我们的模型使用了 BiLSTM+CRF 的深度学习的方法，完全是端到端的识别算法，几乎没有添加任何特征和各种复杂的模板，也达到了 90.75 的 F 值。相比其他模型，我们的模型更加的间接，实用性更强！不过我们的模型还有所欠缺， 为了增加在特定领域的 F 值，我们会在更大的数据集上进行训练以提高 F 值。

（8） 系统界面展示

用户在命名实体识别系统中可以导入文本、输入文本、NER 识别，然后再 result 文本框中就可以看到相关结果。系统还保存了相关日志，以保存系统运行的状态。如下图：

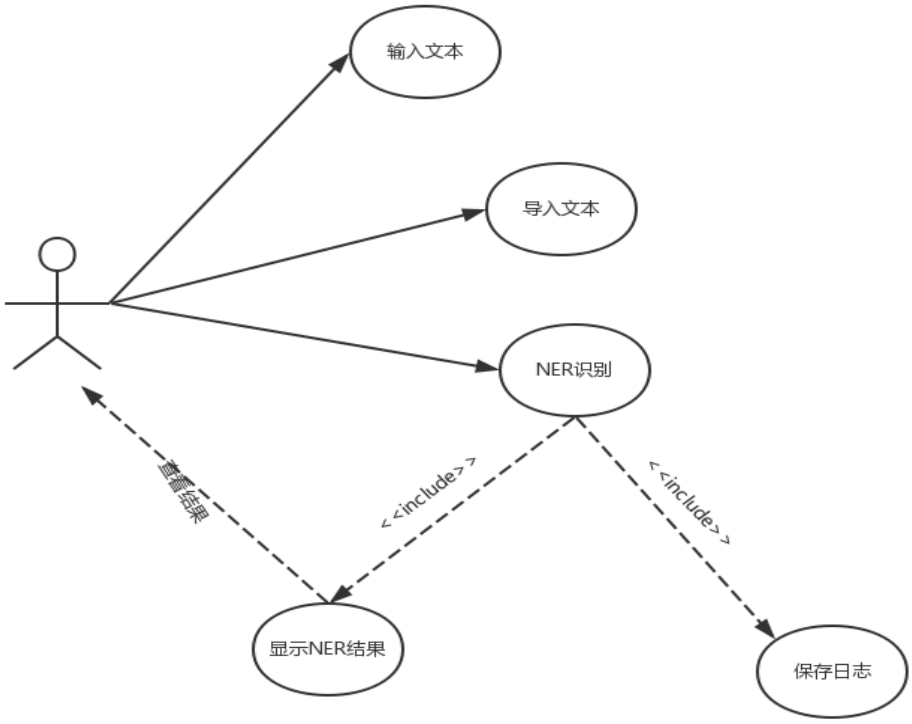

图 2-10 命名实体识别系统用例图

启动系统后，用户可以导入相关文本，然后点击 NER 菜单就可以执行命名实体识别任务，在 result 文本框可以看到最后提取的实体结果！系统运行界面如下：

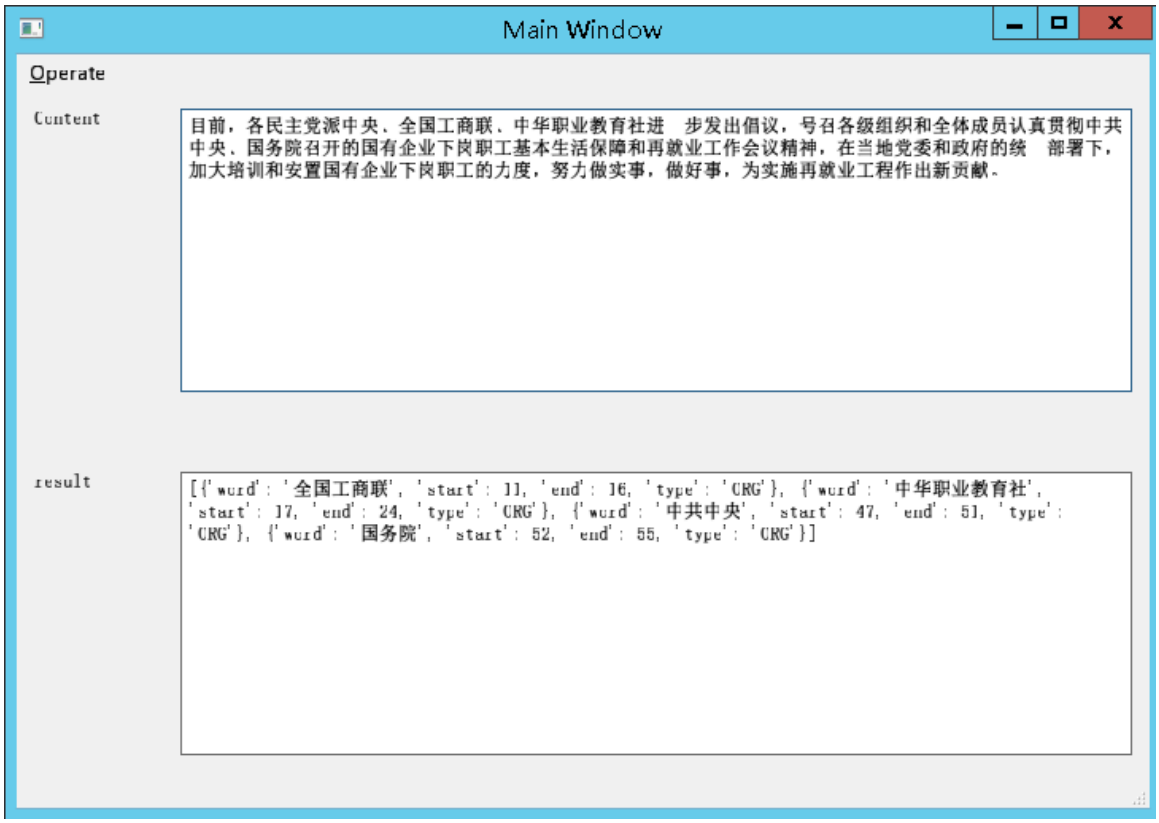

图 2-11 系统运行界面图

2.3 存在的困难与问题

（1）由于深度学习的训练需要很好的硬件设备，特别是 GPU，由于我的 GPU 不是太好，训练特别耗时。一次训练就将近花了五个小时，不太容易调参。

（2） 深度学习模型需要大量的训练数据， 金融领域的实体数据太过稀缺，大部分都需要自己标注，耗时耗力，而且人工标注的数据准确率不太好保证！

（3）深度学习模型可能需要进一步添加部分特征以提高模型在特定领域的通用性，不过怎么提取特征，提取什么样的特征？

2.4 如期完成预定任务的可能性分析

基本上可以实现所有功能，虽然模型目前的 F 值还不错，但是想在特定领域提高 F 值，需要标注大量的金融领域实体识别数据， 还需要进一步做特

征工程的工作。另外需要调整算法结构，让算法更加简洁，更实用，可移植性更强。

2.5 后期工作安排（或进度和计划调整）

后半期的工作进度、时间安排，以及对最初计划的调整情况如下：

（1） 对实现的算法多次进行大规模的数据测试。

（2） 改进已经实现的算法，专门针对金融领域，提高效率和准确率。

（3） 对暂时没有实现的问，比如特征工程，继续学习相关方法，提取比较好的特征来提高正确率 F 值。

（4） 撰写毕业论文，准备答辩。

附件：第 1 次中期检查记录表

<table>
<tr><td>学生姓名</td><td>乐远</td><td>学　　号</td><td>1143710316</td></tr>
<tr><td>专　　业</td><td>软件工程</td><td>中检日期</td><td>2018 年 3 月 31 日</td></tr>
<tr><td>指导教师</td><td>郭勇</td><td>职　　称</td><td>讲师</td></tr>
<tr><td>设计（论文）题目</td><td colspan="3">基于深度学习的命名实体识别系统设计与实现</td></tr>
<tr><td colspan="4">指导教师评语：

该生前期毕业设计工作表现：☐非常突出　☐积极努力　☐较积极　☐一般　☐差；
项目完成工作量约：☐70%以上　☐60%以上　☐50%以上 ☐低于 50%　☐严重不足；
中期报告撰写：内容☐充实　☐较充实　☐一般　☐不够充分；
格式☐规范　☐较规范　☐一般　☐不规范。
☐同意 ☐不同意 ☐报告修改后同意 参加中期检查。
如果保留当时的邮件截图，则删除上述内容，直接粘贴截图即可，不需要导师亲笔签字

签字：</td></tr>
<tr><td colspan="4">答辩记录：

粘贴第 1 次中检意见</td></tr>
<tr><td colspan="2">是否通过中检：　☐通过　☐警告　☐不通过</td><td colspan="2">成绩：☐A　☐B　☐C</td></tr>
<tr><td colspan="2">答辩组长签字：</td><td colspan="2" rowspan="2">答辩组成员：从大表中粘贴</td></tr>
<tr><td colspan="2">答辩秘书签字：</td></tr>
</table>

附件：导师意见邮件截图

注：1.对于亲笔签字的情况不需要此内容；

2.邮件截图中必须含有发件者信息、发件时间、意见等内容。

追思 您好！

中期报告可以了，同意参加中期检查。

****

姓名：郭勇

单位：哈尔滨工业大学软件学院

电话：0451-86417732

邮编：150001

日期：2018-03-30

****