哈 尔 滨 工 业 大 学

国 家 示 范 性 软 件 学 院

本 科 毕 业 设 计(论 文)开 题 报 告

题 目： 基于证券知识图谱构建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	18033066792
校内指导教师	郭 勇、张宇
联 系 方 式	13030000672
开 题 日 期	2017.12.02

哈尔滨工业大学软件学院

目　录

1. 项目来源及开发目的和意义

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。

1.2 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[1]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。

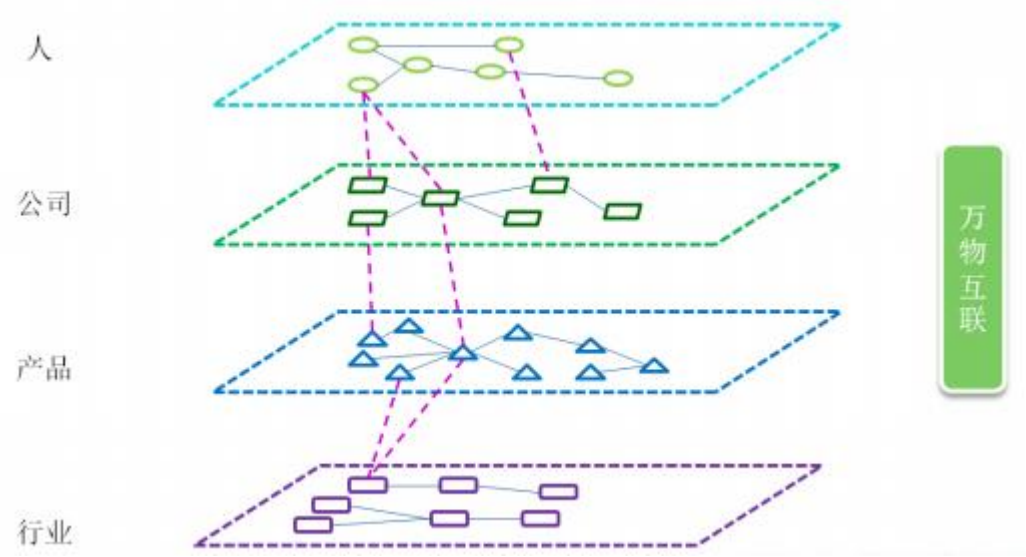

图 1-1 证券知识图谱万物互联图

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。

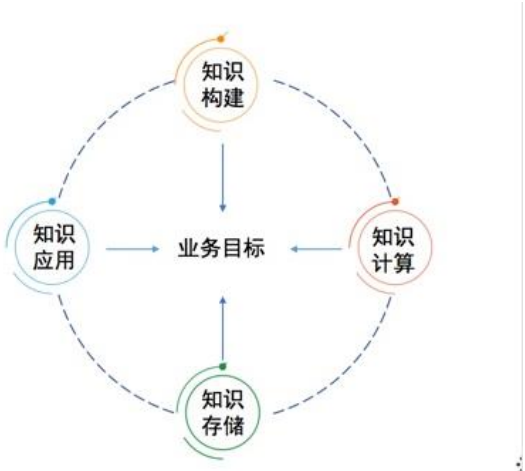

图 1-2 金融知识图谱构建组成部分

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。

2. 国内外相关领域开发及应用现状分析

命名实体识别是一个很古老的自然语言处理任务，1991 年，Lisa F. Rau[2] 进行了一项从文本识别公司及组织名的研究，它才开始进入人们的视线。命名实体这一名词最早是在第 6 届信息理解研讨会（MUC-6）上被提出的。当时的信息理解讨论会主要关注从报纸文章等非结构化文本和信息中，抽取出国防任务等相关内容的结构化信息，主要定义了如识别组织机构名、人民和地点名等相关任务及命名实体识别，并发现其对分析和整理非结构化文本的必要性。在这样趋势和环境下，命名实体识别任务的相关研究得到了重视和快速的发展。

命名实体任务的相关研究是一个漫长而又逐步演变的过程。第一个在该领域研究的论文是 Lisa F. Rau[2] 1991 年在 IEEE 人工智能应用会议上发表的基于启发式搜索和规则匹配的识别公司组织名的的系统。人工撰写的规则一般都是通过对典型的文本中正面和负面的样例，根据人工经验总结而得到的。这样的规则对特定范围的文本的识效果较高，然而人工成本较高，且覆盖率和平均工作效率都非常低。随着统计学习模型的发展以及计算机性能的逐步提高，基于统计学习中有监督学习的命名实体识别相关研究逐渐流行起来。通过自动发现和抽取标注语料库中正面和负面的样例的字、词、频次、位置等特征并建立机器学习模型，对无标注的测试文本进行标注，代替了人工抽取规则的流程。基于统计的命名实体识别方法大幅度的减轻了研究者的人工劳动，并且能够结合研究者人工补充规则更好的提升研究效果，常见的方法有隐马尔科夫模型[3]（HMM, Hidden Markov Model），最大熵马尔科夫模型[4]（ME, Maximum Entropy），支持向量机[5](SVM, Support Vector Machine), 条件随机场[6](CRF, Conditional Random Fields)等模型。

隐马尔可夫模型是一种重要的统计自然语言模型，广泛用于语音识别、词性标注及命名实体识别等领域。2002 年，Zhou[7]利用一个基于 HMM 的组块标注器进行命名实体识别，识别类型包括名字、时间、数字短语。在 MUC-6 和 MUC-7 测试，英文命名实体识别的 F 值分别达到 96.6%和 94.1%。2004 年，Zhao[8]将 HMM 应用到生物医学领域的命名实体识别，同时采用基于词的相似度的平滑方法，当使用大规模未标注的语料时，它可以改善性能，降低数据稀疏问题的影响。基于 HMM 的命名实体识别方法效率较高，但是它融合多种信息的能力不强，同时它不能利用上下文的信息对于复杂的命名实体识别将遇到困难。

最大熵模型利用了信息论中熵的概念，其基本思想是要从全部符合约束条件(通常是给定的某些随机变量的分布)的分布中选出一个使熵值达到最大的分布。1999 年，Borthwich[9]最早将 ME 方法引入到英文命名实体识别问题中，在他的系统中利用了二元特征、词性特征、段落特征、字典信息等。在 MUC-7 上测试，英文命名实体识别总的 F 值为 92.20%。2002 年，Chieu[10]实现了一

个基于最大熵的英文命名实体识别系统，但和其他的系统不同的是，它使用了整个文档的信息，将通常算法中最大化 p(N|S)修正为最大化 p(N|S,Doc),是对最大熵模型的一种扩展。在 MUC-6 和 MUC-7 上测试最好的结果分别达到 93.27%和 87.24%。最大熵方法可以结合更丰富的特征，同时体现出每个特征的重要性。但是它的计算比较复杂而且系统开销比较大。

支持向量机是一种基于统计机器学习理论的模式识别方法，它由 Boster、Guyon、Vapnik 在 COLT92 上首次提出，现在已经在许多领域(生物信息学、文本和手写识别等)得到了成功的应用。在命名实体识别中，SVM 将命名实体识别看做是一个分类问题，但由于 SVM 主要处理二元分类问题，所以在命名实体识别中需要注意。2002 年，Hideki[11]采用讲一个类别和其他所有类别看做两类的思想来解决多分类问题。同时，为了克服 SVM 效率偏低的问题，作者还优化了 SVM 的二次方程 kernel，提高了效率。虽然如此，但是 SVM 的效率还是比较低。

条件随机场[6](CRF, Conditional Random Fields)是由 Lafferty[12]在 2001 年提出的一种基于概率无向图的判别式模型，通过观测序列和待抽取的特征序列，建立对数似然模型进行特征学习。条件随机场模型既解决了隐马尔科夫模型(HMM, Hidden Markov Model)[3]无法根据整句的特征参数优化的缺点，又解决了最大熵模型的标记偏置问题，在解决与序列相关的自然语言处理任务中有很突出的效果，也一直是基于统计的命名实体识别方法中效果表现最好的模型之一，被广泛应用在学术研究和实际应用中。

2003 年，Mccallum[13]在 CoNLL-2003 的命名实体识别的相关任务中，利用了条件随机场模型并取得了相对不错的成绩。与此同时，2005 年 Finkel[14]利用了吉布斯采样（Gibbs sampling）的相关训练方法，并添加了部分非局部的特征，进行了条件随机场模型的训练，并将其应用到命名实体识别任务，也取得了较好的结果。2006 年 Krishnan 等[15]进行了历史研究的总结和整理，对于模型不能够充分的利用到序列标注问题中的局部特征的情况，他们通过对条件随机场的输出进行再次的条件随机场的建模的方法，来进行更好的局部特征提取和利用。而关于如何更好的利用条件随机场模型，进行特征的抽取组合等相关的方向，2008 年张祝玉等[16]进行了充分的对条件随机场的比较实验，利用多种特征组合和抽取，通过多次的比较实验定义相应的特征的贡献度，并选择其中贡献度大的特征，同时还利用了集成模型的方法提升了条件随机场模型的效果。

随着近几年计算机硬件的高速发展，尤其是图形处理器计算能力的提高，深度学习逐渐流行起来，尤其是在图像识别、语音识别已经取得显著的成果，在自然语言处理领域也有突破性的效果。

不过深度学习在自然语言处理的初步探索中，并没有像在其他领域一样，取得突飞猛进的效果。研究者们主键意识到这一问题的核心要点，在与如何对词进行分布式表示，并进行类似传统自然语言处理中的语言模型的模型构建。

2006 年 Bengio 等[17]创造性的提出用向量来表示词，提出了神经网络语言模型。2013 年 Mikolov 等[18]提出 word2vec，创造性地进行了结构上的改进和速度上的优化，使得词向量的训练速度大大加快。除此之外，2010 年 Mikolov 等[19]在 Bengio 等的基础上，提出了循环神经网络语言模型，能够更好的考虑到上下文信息，在处理序列数据，如文本数据等具有很大的优势，效果也得到了大幅度的提升。2014 年，在词向量的基础上，研究者又提出了句向量[20]、文档向量[20]，这些方法在实际应用中都发挥了其独有的效果。

另一方面特定结构的神经网路也影响着自然语言处理的发展。RNN 首先被提出来训练语言模型，它可以带有上下文信息，更符合文本信息的语言特征，不过其缺点也逐渐被发现。由于深度神经网络传递残差时需要多次求导，如果网络的层数过高，经常会出现梯度消失的问题；除此之外，RNN 网络虽然能够保留历史信息，但是由于距离当前输入距离较远的历史信息，会在多次求导不断被稀疏，使得 RNN 网络更多的是保留距离当前输入较近的信息，并不能真正保留所有历史信息，这在句法分析、关系抽取等任务中会有较大的影响。

对于 RNN 模型的如上问题如何进行改进，研究者们进行了许多方面的尝试。1997 年 Hochreiter 等[21]创造性的提出了一种基于 RNN 模型的改进模型：长短期存储单元(Long-Short Term Memory, LSTM)，之后的 Gers 等[22]人、Gravers 等[23]人又在 LSTM 模型的基础上继续努力改进，相比于 RNN 模型仅仅简单使用前一位置的隐含层作为历史信息，不仅保存的信息较为简单，还容易造成梯度消失这一现状，LSTM 模型使用了多个类似于 RNN 隐含层的单元，即门（gate）来控制历史信息和输入信息如何进行输入、输出和更新。除此之外，LSTM 的关键在于使用了一个名为存储单元(Memory Call)代替了传统 RNN 模型的常规神经元，并建立了门机制。其中输入门(input gate)决定输入的信息如何流入到模型中，输出门用于将真正的需要的信息流入到隐含层中，遗忘门则能够控制存储单元何时遗忘，遗忘多少历史信息。相比于 RNN 中仅有一个隐含层来处理历史数据的情况，LSTM 通过三个类似的门单元进行数据控制，能够更有效的保存更有价值的历史信息。

LSTM 模型的提出，立刻吸引了诸多自然语言处理研究者的目光，2014 年 Bahdanau 等[24]将 LSTM 模型组合成端到端的结构，并加入了注意力（attension）机制，显著提高了机器翻译的研究结果。2015 年 Dyer[25]等在依存句法分析任务中使用了 LSTM 模型，在中文和英文两个测试集上都取得了显著的成果。Xu 等[26]利用文本的最短依赖路径信息进行关系分类问题的 LSTM 模型的建模，并加入了词向量与语言学特征向量相结合的方式，在 SemEval 2010 数据集上获得了 83.7%的 F1 值的突出效果。2016 年 Lample 等[27]在命名实体识别任务上分别使用了带有 CRF 转移层的 LSTM 和栈式 LSTM，在多个语种的命名实体识别测试语料上都取得了最好的成绩。

综上，传统的 CRF 和现在的 LSTM 等在命名实体识别任务上都取得了不

错的效果。相比于常规领域的文本，金融证券领域的命名实体识别更加复杂，而且标注数据稀缺，那么如何将这些技术运用到实际项目，怎么在这个实用性更强的领域取得更好的效果将是本项目的一个重要目标。

3. 需求分析及总体设计方案

3.1 主要开发内容

如图 3-1，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。

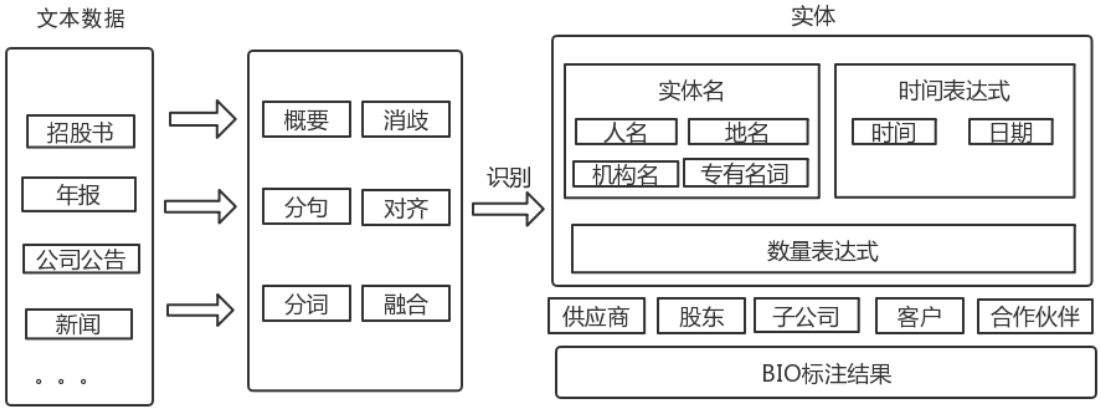

图 3-1 命名实体识别系统主要功能示意图

3.2 需求分析

对于一个命名实体识别算法来说，它必须要在给定的文本中识别出实体名、时间表达式、数量表达式，准确率需要达到一定的要求，F 值也必须满足一定的要求。由于该系统的特殊性，金融领域的招股书、年报、公司报告、新闻等数据既包含结构化数据又包含非结构化数据，因此该系统需要更强的数据预处理能力，比如需要正文提取，形成概要，过滤掉无关信息。另外由于金融领域的命名实体识别更关注供应商、股东、公司、客户、合作伙伴等等相关实体，实体识别更加特殊复杂，容易存在歧义，实体别称多，因此需要消歧、对齐、融合等等。

3.2.1 系统功能需求

本命名实体识别算法系统主要包含正文提取与概要、分句、分词、实体识别、消歧、对齐、融合等 7 个功能模块，文本提取与概要需要从从招股书、年报、公司报告、新闻等结构化和非结构化数据中提取出正文和概要，分句、分词需要能够得到更便于算法处理的文本数据，实体识别需要识别出给定文本中的命名实体，尤其是金融证券领域所关心的公司名、人名、合作伙伴、供应商等等，而消歧、对齐、融合需要对得到的 BIO 标注集合进行的后处理操作。

如表 3-1 所示，该命名实体识别系统几个模块详细功能点及需求列出如下。

表 3-1 命名实体识别需求分析表

模块	需求描述
正文提取与概要	1. 从招股书、年报、公司报告、新闻等结构化和非结构化数据中提取出正文和概要
分句	1. 将文本分成句子集合
分词	1. 将句子分词
实体识别	1. 使用命名实体识别算法识别出实体名、时间表达式、数量表达式 2. 输出相应的 BIO 标注结果
消歧	1. 消除由于文本歧义带来的不正确的实体识别结果
对齐	1. 将异构数据源知识库中的各个实体，找出属于现实世界中的同一实体
融合	1. 对不同数据源中的实体信息进行整合

下面详细介绍本命名实体识别算法的重点部分：实体识别。

命名实体识别（Named Entity Recognition，简称 NER），又称作“专名识别”，是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等。 一般来说，命名实体识别的任务就是识别出待处理文本中 三大类（实体类、时间类和数字类）、七小类（人名、机构名、地名、时间、日期、货币和百分比）命名实体。

由于金融证券领域的特殊性，主要关注人名、公司名、法院、银行、交易所、工商局、股票名、股权名、基金名等等，因此要着重对这些实体进行识别，这也是本系统最为主要的功能需求。

对于识别出的结果，我们使用 BIO 标注集进行标注。其中 B 表示实体开始，I 表示实体内部，O 表示非实体类型。为了识别更多类型的实体，另外还增加一些新的标签：

- B-PERS，B-DATE，B-LOC…… 一个人或者日期实体的开始
- I-PERS, I-DATE, I-LOC …… 一个人或者日期实体的内部
- O 非命名实体

3.2.2 系统非功能需求

本命名实体识别算法的非功能需求主要包括性能需求、易用性需求、扩展性需求、正确性需求，健壮性需求。对于各个非功能需求，其要求如下所述。

（1）**性能需求** 能够处理一定复杂度的数据，能够在可以接受的时间内分析出命名实体结果，不能占据太多内存（不超过 2G）。

（2）**扩展性需求** 接口设计符合开闭原则，允许在不能的功能上采取不同的方法而不影响系统的运行。

（3）**正确性需求** 命名实体识别的正确率必须在 90%以上，F 值在 80 以上。不能有太多歧义的结果。

（4）**健壮性需求** 能够捕捉到数据格式等错误，不会因为突发事件导致数据丢失或者出现差错。

3.3 总体设计方案

3.3.1 系统功能设计

本命名实体识别系统主要包含预处理、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 3-2 所示。

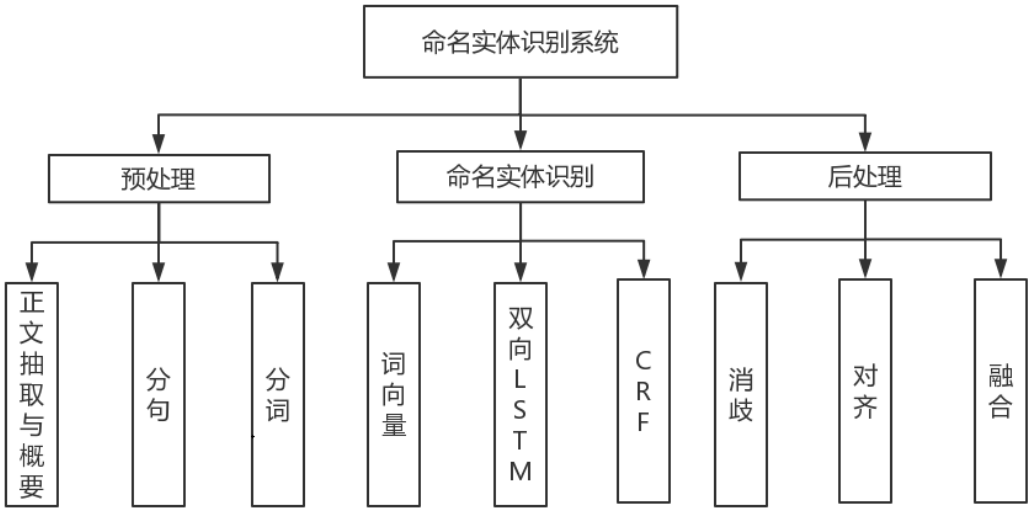

图 3-2 命名实体识别系统功能结构图

3.3.2 实体识别系统概要设计

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，该命名实体识别系统程序流程图如下：

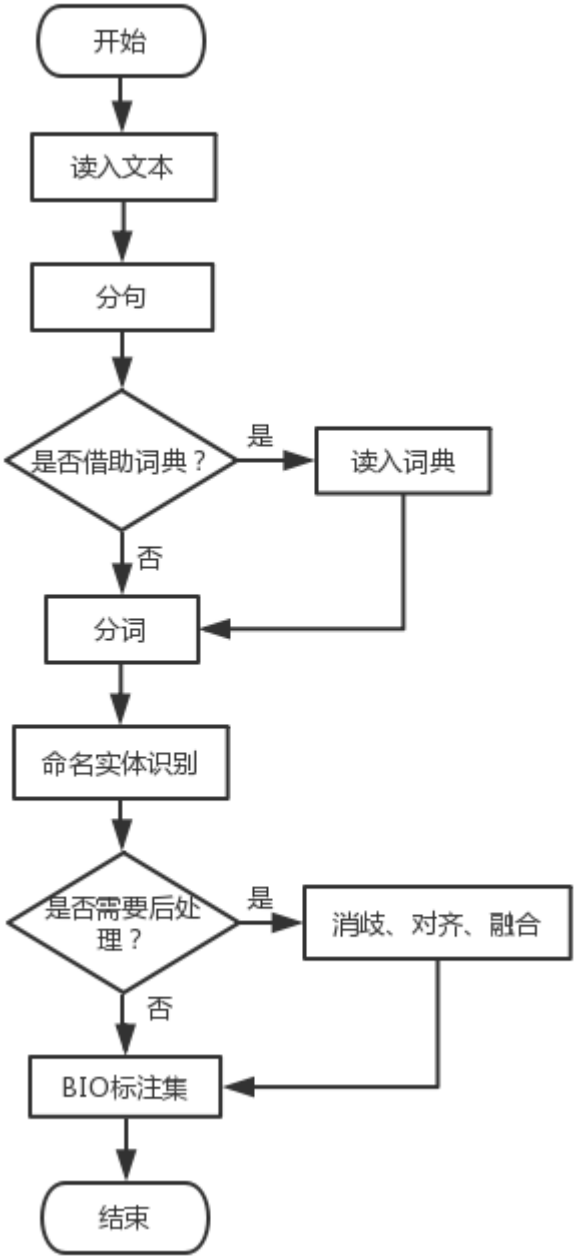

图 3-3 命名实体识别系统流程图

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，这个涉及到文本信息的过滤等技术。然后分句、分词，分词可以使用字典，这个可以使用开源的 NLTK、结巴分词等工具，然后进行实体识别得到 BIO 标注结果，这也是最为核心的一部分。最后判断是否进行消歧、对齐、融合，避免信息的冗余。

3.3.3 实体识别算法设计

目前中文领域的命名实体识别最好的几个算法之一是 CRF++并结合特征模板，不过这个方法太过复杂，不易于迁移到新的领域。随着深度学习在自然语言领域的发展，2016 年 Lample[27]等人提出了一种 LSTM+CRF 的方法，能够在四种语言的命名实体识别任务上能够达到不错的正确率和 F 值，并且在

标注数据稀缺的情况下仍然能够达到不错的效果。因此，我进一步将该方法应用到金融领域的中文命名实体识别任务上，并更改了部分神经网络结构以提高正确率和 F 值，提高命名实体识别的效率。

下面主要介绍利用不同的架构层与双向 LSTM 单元的组合，并通过模型构建时的各种细节和训练技巧，建立一个基于 LSTM 单元的完整的命名实体识别的深度学习模型。

（一）总体架构

图 3-4 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。此处可以利用 word2vec、Glove Vectors 等。

（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 BLSTM 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。

（二）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：

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

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

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

$$h_t=o_t\odot\tanh(c_t) \quad (4)$$

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。

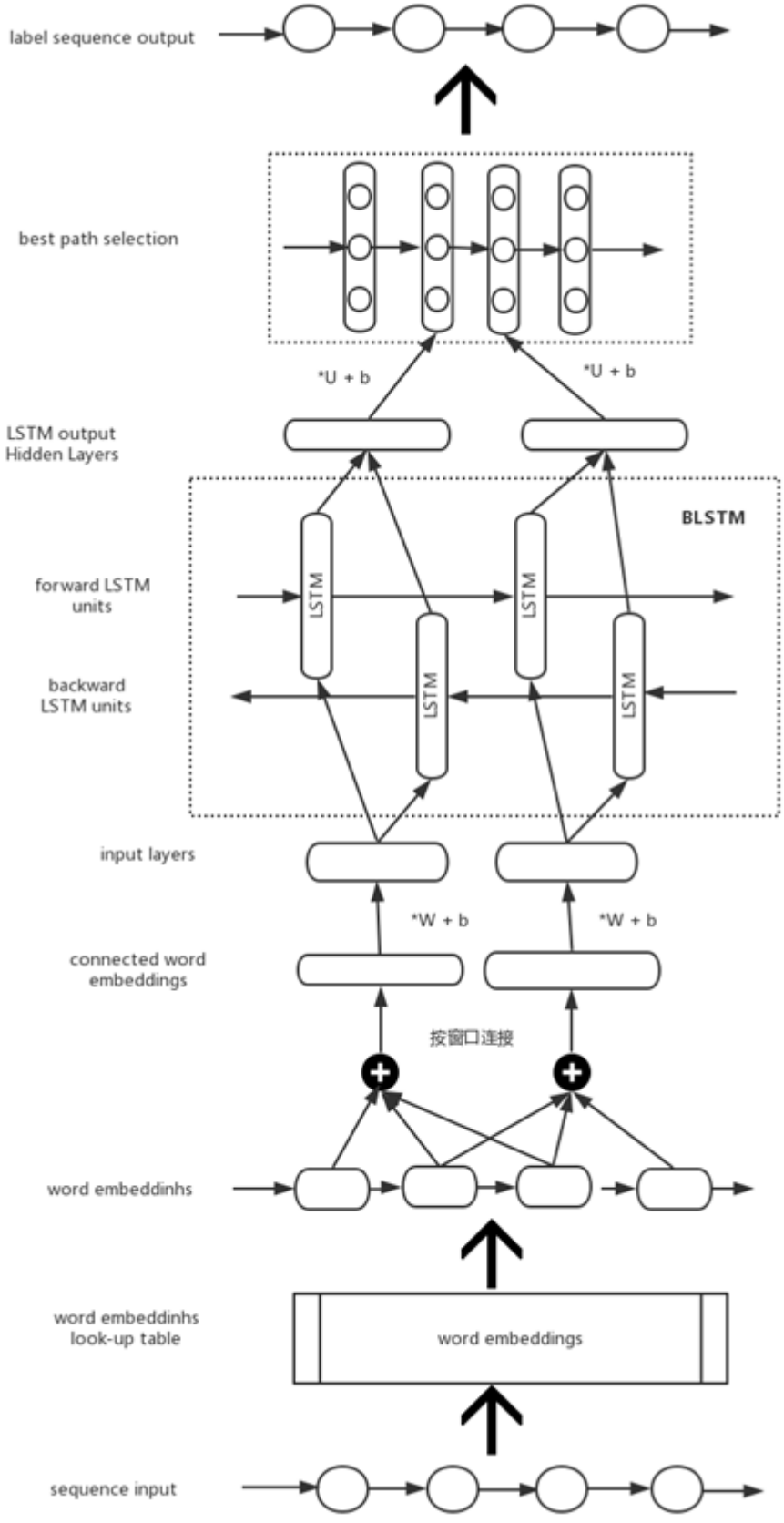

图 3-4 基于 LSTM 的 NER 模型整体框架

(三) 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 3-4 中“按窗口连接”可以看到。

（四）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。

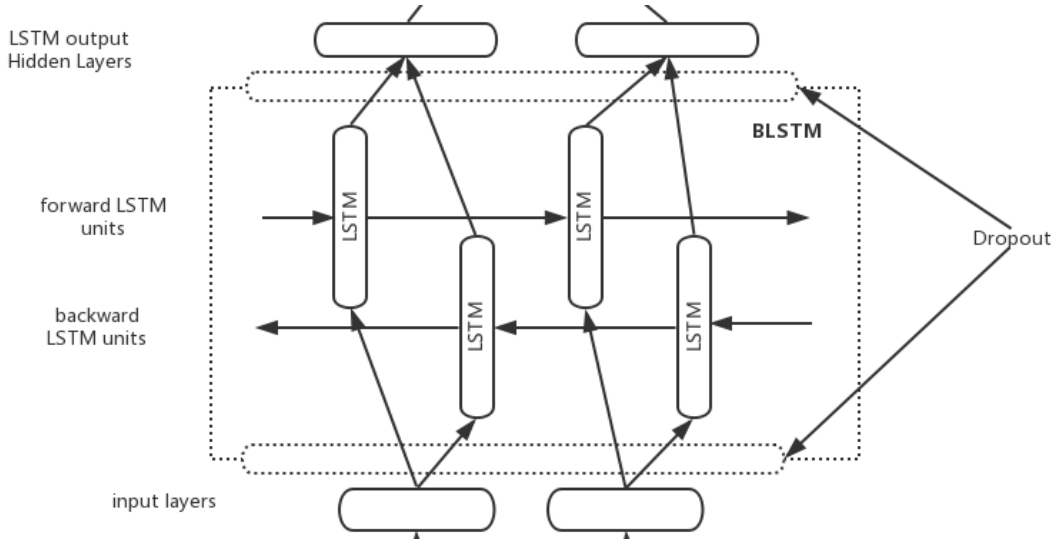

图 3-5 NER Dropout 机制加入位置

（五）引入转移代价的代价计算

借鉴于 Collobert[28]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。参照传统的如 CRF、HMM 等广泛应用于 NER 的模型，在模型的训练的过程中首先使用极大似然法进行优化，然后再利用这个矩阵计算出标注正确结果的路径概率，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：

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

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：

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

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：

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

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：

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

3.3.4 系统功能模块与架构

如图 3-6，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。

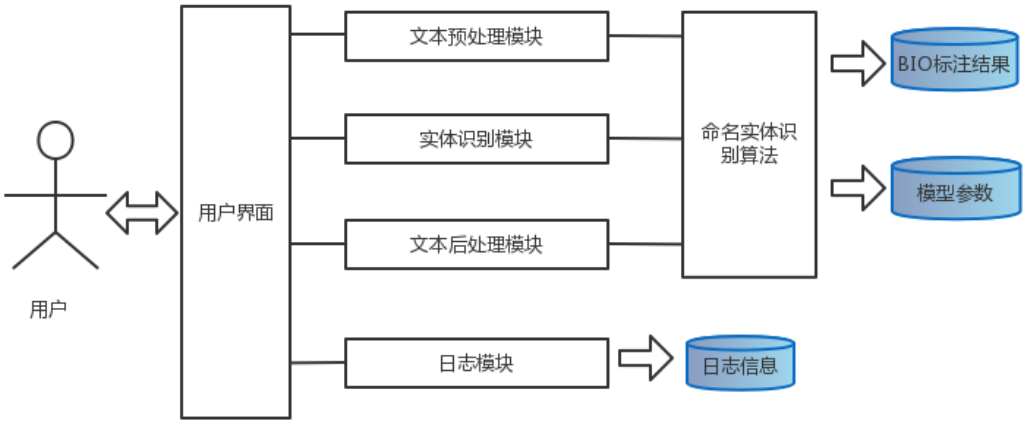

图 3-6 命名实体识别系统架构图

4. 开发环境和开发工具

4.1 开发语言

本系统主要使用 python 语言开发。

4.2 开发工具

本系统使用 Theano 和 Tensorflow 作为开发的深度学习框架，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 4-1 所示，本系统的开发使用到了这些开发工具：

表 4-1 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

4.3 开发环境

操作系统：Windows 10

处理器：Intel Core i5，2.50GHz 主频

内存：8GB 1000 MHz DDR3

程序运行环境：Windows 7 或更高版本的 Windows 操作系统

5. 项目进度安排、预期达到的目标

5.1 进度安排

项目进度及毕业设计（论文）工作安排见表 5-1。

表 5-1 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	在进行
2017.11.11	2017.11.20	完成开题报告，答辩 PPT	未完成
2017.11.21	2017.11.30	收集相关数据，标注数据集	未完成
2017.12.01	2018.01.31	实现深度学习算法，训练，调参	未完成
2018.02.01	2018.03.10	初步完成 NER 的基本识别功能	未完成
2018.03.11	2018.03.20	准备中期答辩	未完成
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

5.2 预期达到的目标

5.2.1 总体目标：

（1）因为该命名实体识别算法系统将会被用于构建金融领域的知识图谱，因此《证券金融知识图谱》项目的目标就是该项目的目标。

（2）本系统作为一个算法类项目，自己借鉴了相关论文的所使用的 LSTM+CRF 深度学习模型，并在该基础之上更改了部分结构，该模型必须保证能够具有更高的正确率和 F 值。

（3）相比于一些开源的命名实体识别工具，这个能够满足特定领域的命名实体识别任务，更具有实用性。

（4）命名实体识别是自然语言处理领域一个应用很广泛的任务，为以后研究相关的课题或领域打下坚实的基础。

5.2.2 功能目标：

（1）能够对输入的招股书、研报、公告、新闻等进行正文提取，文本概要。

（2）对输入的句子能够进行训练，并得到相关的词向量的语义表示。

（3）对输入的句子能够进行较高的正确率进行标注，并给出 BIO 的标注结果。

5.2.3 性能目标：

（1）对招股书、研报、公告、新闻等数据的处理要正确，不能遗漏、弄错重要信息，识别实体的正确率必须在 90%以上。

（2）对给定的文本进行命名实体识别不能耗费太多时间，如 10000 字的文章识别实体不能超过 1 分钟。

6. 完成项目所需的条件和经费

6.1 已具备的条件

（1）开发用的个人笔记本电脑和公司台式电脑，操作系统为 Windows 10；

（2）python2.7 编程环境，PyCharm 2017 开发平台；

（3）机器学习、深度学习相关理论知识已有一部分基础；

（4）python、Theano 一般语法已经掌握，能够实现并训练部分简单的神经网路；

（5）机器学习相关项目的实际经验，如公告分类项目；

（6）自然语言处理有一定基础；

（7）对深度神经网络 RNN、LSTM 的算法理论有一定的了解；

（8）github 版本控制使用基础；

6.2 需要的条件和经费

（1）周志华《机器学习》;

（2）深度学习相关课程和书籍、论文；

（3）命名实体识别的相关论文；

本项目所有书籍经费以及设备均由本人和公司共同承担。

7. 预见的困难及应对措施

本项目开发过程中，可以预见的困难及应对措施如下：

（1）对深度学习中 RNN、LSTM 的模型机制理解还不够深刻，虽然其在英文的命名实体识别任务中效果很不错，可是怎么将该方法借鉴到中文的命名实体识别的任务上还很难，而且能不能达到很高的正确率还不太确定。

应对措施：广泛阅读该领域的相关论文，先不着急实现这个命名实体识别算法，先学习大牛们怎么运用 LSTM 等神经网络模型来进行命名实体识别，然后再学习一下怎么用神经网络来进行中文的命名实体识别，重点注意他们的网络结构、训练思想方法与技巧，最后在来思考实现自己的模型方法。

（2）对 python、Theano 和 Tensorflow 框架的理解运用还不够熟练，实现多层的神经网路结构模型的可能会比较的困难，而且神经网络模型参数巨多，怎么调参优化不仅耗费时间，还有很多的技巧与难点。

应对措施：积极地阅读大牛们的代码，学学他们怎么用 python 相关框架实现深度学习网络模型，学习他们的训练方法，逐步地积累经验。

（3）金融领域的命名实体识别训练数据稀缺，大量的标注工作需要做，而且招股书、研报、公告、新闻等数据结构冗杂，怎么提取出主要信息文本是一个很大的问题。

应对措施：寻找一些已经开源的标注好的金融领域相关的数据集，并结合深交所特定应用领域再来标注一些数据集，最后结合这些数据集来进行训练测试。

（4）Theno 和 Tensorflow 是一个比较底层的深度学习框架，训练效率相对较低，对 GPU 的配置要求，特别是中文语料库比加大的时候一个实验能跑一周，太耗费时间而使得调整参数的时间会少很多。

应对措施：学习使用更高级的框架如 keras 等，并运用多线程编程的技术加快训练速度，节省开发时间。

参考文献

[1] Wenliang Chen, Chenyu Wang, Bing Xiao, Weining Qian and Aoying Zhou. On Statistical Characteristics of Real-life Knowledge Graphs. In Big Data Benchmarks, Performance Optimization, and Emerging Hardware – 6th Workshop, BPOE 2015, pages 37-49,2015.

[2] Rau L F. Extracting company names from text[C]//Artificial Intelligence Applications, 1991. Proceedings., Seventh IEEE Conference on. IEEE, 1991, 1:29-32.

[3] Bikel D M, Miller S, Schwartz R, et al. Nymble: a high-performance learning name-finder[C]

[4] Proceedings of the fifth conference on Applied natural language processing. Association for Computational Linguistics, 1997: 194-201.

[5] Klinger, Roman, and Katrin Tomanek. Classical probabilistic models and conditional random fields. TU, Algorithm Engineering, 2007

[6] Asahara M, Matsumoto Y. Japanese named entity extraction with redundant morphological analysis[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003: 8-15.

[7] Cuodong Zhou，Jian Su.Named Entiyt Recognition using an HMM-based Chunk Tagger. ACL，PhiladelPhia，USA，2002:473 一 480

[8] ShaojunZhao. Named Entiyt Recognition in Biomedical Texts using an HMM Model. JNLPB，2004

[9] A Borthwick Mxaimum EntroPy APProach to Named Entity Recognition.PhD Dissertation，New York University，1999:18-25.

[10]Hai Leong Chieu，Hwee Tou Ng.Named Entity Recognition:A Mxaimum EntroPy APProach Using Global Inofrmation.COLING, TaiPei，Taiwan，2002.

[11]Hideki Isozaki，Hideto Kazwaa. Efficient Support Vector Classifiers for Named Entity Recognition.COLNIq2002:953-959.

[12]Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the eighteenth international conference on machine learning, ICML. 2001, 1: 282-289.

[13]Andrew Mccallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In HLT-NAACL, pages 188–191.

[14]Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. "Incorporating non-local information into information extraction systems by gibbs sampling." Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics, 2005:363–370.

[15]Krishnan, Vijay, and Christopher D. Manning. "An effective two-stage model for exploiting non-local dependencies in named entity recognition." Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006:1121–1128.

[16]张祝玉, 任飞亮, 朱靖波. "基于条件随机场的中文命名实体识别特征比较研究 [C]." 见: 第 4 届全国信息检索与内容安全学术会议论文集. 2008.

[17]Bengio Y, Schwenk H, Senécal J S, et al. Neural probabilistic language models[M]//Innovations in Machine Learning. Springer Berlin Heidelberg, 2006:137-186

[18]Tomas Mikolov, Kai Chen, Greg Corrado, and Jef-frey Dean. 2013a. Efficient Estimation of Word RepresentationsinVectorSpace. In ICLR Work-shop Papers.

[19]Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]//INTERSPEECH. 2010, 2: 3.

[20]Le Q V, Mikolov T. Distributed representations of sentences and documents[J]. ar Xiv preprint ar Xiv:1405.4053, 2014.

[21]Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[22]Gers F A, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM[J]. Neural computation, 2000, 12(10): 2451-2471.

[23]Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5): 602-610.

[24]Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. ar Xiv preprint ar Xiv:1409.0473, 2014.

[25]Dyer C, Ballesteros M, Ling W, et al. Transition-based dependency parsing with stack long short-term memory[J]. ar Xiv preprint ar Xiv:1505.08075, 2015.

[26]Xu Y, Mou L, Li G, et al. Classifying relations via long short term memory networks along shortest dependency paths[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing (to

appear). 2015.

[27]Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[J]. ar Xiv preprint ar Xiv:1603.01360, 2016.

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

附件 1：哈尔滨工业大学毕业设计（论文）任务书

姓 名：乐 远	院（系）：软件学院
专 业：软件工程	学 号：1143710316
任务起止日期： 2017 年 07 月 17 日 至 2018 年 07 月 01 日	

毕业设计（论文）题目：

基于证券知识图谱构建的命名实体识别系统设计与实现

立题的目的和意义：

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。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，从而达到人、公司、产品、行业的“万物互联”，从而提高行业信息利用的精准度和可信度，以及广度。而构建知识图谱的基础核心技术之一就是命名实体识别算法设计，为了就是更高效、准确地从各类结构化和非结构化的文本数据中识别出金融领域关心的人名、公司名、银行、法院、交易所等等实体，从而在此基础上构建更准确、实用、功能更强大的知识语义网络，为金融领域的相关应用提供可靠的信息源。

技术要求和主要内容：

需要灵活使用 Theano、Tensorflow 等深度学习框架，并且需要熟练掌握 python 基本语法知识，掌握常见 numpy、scicpy、matplotlib 等库，深度理解机器学习特别是深度学习相关的理论知识，能够熟练地运用 python 实现这些深度学习模型算法，并且能够调参，能够分析结果并根据结果能够对模型进行优化调整，能够针对具体问题分析原因并能够给出解决问题的方案。

另外英文文献的阅读能力、金融领域相关文档的阅读能力也很重要。

进度安排：

2017 年 07 月 25 日—2017 年 09 月 01 日 跟随导师了解项目，毕设选题
2017 年 09 月 02 日—2017 年 09 月 30 日 完成项目需求与可行性分析
2017 年 10 月 01 日—2017 年 10 月 12 日 完成概要设计
2017 年 10 月 13 日—2017 年 10 月 31 日 阅读相关论文，调研领域算法
2017 年 11 月 01 日—2017 年 11 月 10 日 撰写开题报告，准备开题答辩
2017 年 11 月 10 日—2017 年 11 月 20 日 完成开题报告，答辩 PPT
2017 年 11 月 21 日—2017 年 11 月 30 日 收集相关数据，标注数据集
2017 年 12 月 01 日—2018 年 01 月 31 日 实现深度学习算法，训练，调参
2018 年 02 月 01 日—2018 年 03 月 10 日 初步完成 NER 的基本识别功能
2018 年 03 月 11 日—2018 年 03 月 20 日 准备中期答辩
2018 年 03 月 21 日—2018 年 04 月 21 日 继续调整优化算法，调整模型结构
2018 年 04 月 22 日—2018 年 05 月 22 日 完善算法的设计细节
2018 年 05 月 23 日—2018 年 06 月 15 日 用多组大规模数据进行训练，测试
2018 年 06 月 16 日—2018 年 06 月 20 日 完善最后的一些细节
2018 年 06 月 21 日—2018 年 07 月 01 日 撰写、修改论文，参加毕业答辩

同组设计者及分工：

无

指导教师意见：

签 名：

年 月 日

教研室主任意见：

签 名：

年 月 日

附件 2：本科毕业设计(论文)开题检查意见表

<table>
<tr><td>基地指导教师意见（需写具体内容）</td></tr>
<tr><td>签　字：
年　月　日</td></tr>
<tr><td>校内指导教师意见（需写具体内容）</td></tr>
<tr><td>乐远 您好！
报告可以了，同意参加开题检查！

此致！

2017-11-30
****
发件人：郭勇
单　位： 哈尔滨工业大学 计算机科学与技术学院
电　话： 86417732-808
guoy@hit.edu.cn
****
签　字：
年　月　日</td></tr>
<tr><td>开题检查小组意见</td></tr>
<tr><td>结论：◎ 通过
◎ 警告
◎ 不通过

具体意见：</td><td>评委签字：

年　月　日</td></tr>
</table>