本科生毕业论文（设计）

毕业论文（设计）题目：

涉恐文本中的命名实体识别方法研究

部　　院　系：　信息科学与技术学院

专　　　业：　计算机科学与技术

学　　　号：　201411212041

学 生 姓 名：　朱彦丞

指 导 教 师：　别荣芳

指导教师职称：　教授

指导教师单位：　北京师范大学信息科学与技术学院

2018 年 5 月 16 日

北京师范大学本科生毕业论文（设计）诚信承诺书

本人郑重声明： 所呈交的毕业论文（设计），是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

本人签名： 年 月 日

北京师范大学本科生毕业论文（设计）使用授权书

本人完全了解北京师范大学有关收集、保留和使用毕业论文（设计）的规定，即：本科生毕业论文（设计）工作的知识产权单位属北京师范大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许毕业论文（设计）被查阅和借阅；学校可以公布毕业论文（设计）的全部或部分内容，可以采用影印、缩印或扫描等复制手段保存、汇编毕业论文（设计）。保密的毕业论文（设计）在解密后遵守此规定。

本论文（是、否）保密论文。

保密论文在_____年_____月解密后适用本授权书。

本人签名： 年 月 日

导师签字： 年 月 日

涉恐文本中的命名实体识别方法研究

摘　要

近年来，恐怖袭击事件在全球范围内愈演愈烈，不仅给人们的生命和财产构成了严重的威胁，同时还给人们的精神带来了难以承受的负担。为应对这一挑战，从涉恐文本中挖掘出人名、地名、组织机构名的实体能够为指定反恐行动计划的部门提供战略指导。而通过比较不同算法和框架，可以深入理解命名实体特点和为涉恐文本中的命名实体抽取提供更为快速准确的方案。论文的具体工作如下：

首先，本文分析基于关键词索引从百度搜索中获得涉恐事件的文本内容，然后对基于两大命名实体识别框架（玻森 NLP 和哈工大语言云 LTP）和两大算法（条件随机场算法以及神经网络算法）来对已标注的中文命名实体识别语料进行评测分析，提出了使用字词级别的卷积神经网络和循环神经网络相拼接的方式来优化命名实体识别算法，该模型既利用了预测词的窗口上下文信息又使用了句子级别的语义信息，结果表明这些新增的信息对于识别的准确率和召回率有着显著的提升。论文还使用 LTP 的 API 来自动化抽取涉恐文本中的命名实体从而建立涉恐事件的文本检索列表。

关键词：涉恐文本， 命名实体识别， 玻森 NLP， 哈工大 LTP， 条件随机场算法， 神经网络算法

Research on Named Entity Recognition Methods in Terrorism Text

ABSTRACT

Recently, terrorist attacks have intensified on a global scale, which not only poses a serious threat to people's lives and property, but also brings unbearable burdens to people's spirits. To address this challenge, the entity, including person, place and organization name, which is excavated from the terror-related texts can provide strategic guidance for the departments that designated anti-terrorism action plans. By comparing different algorithms and frameworks, we can learn more about the characteristics of named entities. And this can also provide a faster and more accurate solution for the extraction of named entity information in terror-related texts.

Firstly, this article uses the keyword index to obtain the contents of the terror-related text from Baidu search and then I use annotated corpus of Chinese NER to compare the two major NER framework (Boson NLP and HIT LTP) and two algorithms (Conditional Random Fields and Neural Networks) training results. Then I put forward a method by concatenating the character-level and word-level's Recurrent Neural Network and Convolutional Neural Network to optimize the Named Entity Recognition Algorithm. This model not only uses the window context of the predicted word but also uses the semantic information of the sentence. With those additional information, the performance of recognition's precision and recall has been greatly improved. Finally, LTP is used to

automatically extract the named entities from terror-related texts to build terrorism retrieval lists.

KEY WORDS： Terror-related texts, Named entity recognition, Boson NLP, HIT LTP, Conditional Random Fields, Neural Networks

目 录

1 引言

1.1 涉恐文本概述

涉恐从广义上是指涉及恐怖袭击的一切活动，是以个人或者某种组织发起的为制造社会恐慌、危害公共安全等对一国政府或者社会公众使用暴力、破坏、恐吓等相要挟，使人丧失安全感以及引起社会秩序混乱、造成财产损失和人员伤亡等严重后果的有预谋的犯罪行为。自美国 911 事件以来，恐怖袭击事件在全球范围内愈演愈热烈，各国政府都在建立反恐情报来收集恐怖分子及组织的事件和行为来预防和打击恐怖主义。例如美国主导的“全球恐怖主义数据库（GTD）”就详细的记录了 1970-2016 年间全球发生的恐怖袭击行动，其中也包括了对于中国的记载，该数据库的建立有助于追踪恐怖主义的源头和谱系，也是了解恐怖主义发展趋势的重要来源。

涉恐文本是了解各大恐怖主义组织发展的重要情报也是国家政府安全部门可以快速反应和处理潜在威胁的重要资料来源。其文本来源包括公开途径（如新闻媒体报道，网站的披露等）和非公开途径（如政府的公安部门情报等），其中网络的快速便捷的发展使得大众更容易从各大门户网站上获得最新的资讯和新闻报道信息。在涉恐的新闻文稿中往往会涉及到恐怖袭击事件的时间、地点、人物、袭击类型和袭击过程等，犯罪分子往往都有着相似的作案手法或者逃离路线，这些公开信息的搜集和处理有助于建立一套完整的情报系统从而掌握恐怖活动在时间和空间维度上的发展，为安全决策作出及时的反应。

1.2 命名实体识别概述

命名实体识别(Named Entity Recognition)是自然语言处理中一个常见的经典问题，是指从文本中识别出命名性的指称项，无论在搜索、推荐系统还是其他方面都得到了广泛的应用。狭义上是识别出人名(PER)、地名(LOC)和组织机构名(ORG)这三类命名实体，本论文采用的是三类别，但广义上还包括了除上述的三种外的日期、时间、百分数、货币共七种命名实体。

图 1-1 命名实体示意图

实体的标注体系存在 BIO 以及 BIESO 这两种不同的方法，本论文中使用的是 BIO 标注体系。

表 1-1 BIESO 标注体系

标准符号	O	S	B	I	E
含义	不是 NE	单独构成 NE	NE 的开始	NE 的中间	NE 的结尾

命名实体识别算法作为 NLP 领域中的基础任务得到了广泛的研究，但汉语作为象形文字相比其他拼音文字（如英语及其他拉丁语系）来说具有更多的挑战性：

1） 词语界线标志：中文没有明确的可以分割开词语的界限，如空格或者是字母的大小写指示。

2） 上下文语境：中文的词语中往往存在着一词多义，有些词语脱离了上下文之后便无法判断，而且即便是同一个词语，在不同的语境中可能分属于不同的实体类型。

3） 实体的嵌套：在组织机构名中最为常见，如上文示例中的“北京师范大学”的组织机构名中就嵌套了“北京”这一地名。

4） 缩写简述表达：如“北师大”也可以指代“北京师范大学”，即也应该被识别为组织机构名。

因此如何就中文命名实体在涉恐文本的抽取以及对于命名实体识别算法的评测和应用成为了本课题的研究重点。

1.3 国内外研究现状

目前的命名实体任务都聚焦在商业化的领域中如商品名称的搜索检索上，未涉及到反恐这一专业领域上的识别和应用。但全球恐怖主义数据库的发展促使国内学者考虑和研究中国反恐数据库的建设，例如周松青就对 GTD 做了全面的分析和从理论上提出了中国反恐数据库的战略假设[1]；李本先基于我国当前的反恐指挥体系提出了反恐情报预警体系的基本框架[2]；魏静利用多模块的贝叶斯网络提出了对恐怖袭击的威胁评估的计算方法[3]；孙菲菲则使用了旋转森林集成学习的方法对涉恐实体进行识别和数据挖掘[4]。上述文献大都聚焦于理论上的架构，少部分对从提取命名实体到风险评估的决策分析进行了算法上的分析和研究。

命名实体识别作为建立反恐情报中最为基础和重要的一环，为最后的恐怖袭击的决策评估和研究发展趋势都提供了极为重要的参考。因此本课题就如何提升识别的准确率做了如下的研究：从语言学角度分析，命名实体识别属于中文分词中未登录词识别的任务，既可以采用基于规则匹配方法[8]的识别方法，也可以作为序列化结构的标注采用随机场算法[15]或者是隐马尔可夫[16]算法。

其中基于规则匹配的方法是指将文本与提前手工构造的规则进行模式匹配来识别出命名实体，如“大学”就可以作为组织结构的结尾，“老师”就可以作为人名的下文，除此之外还可以包含词性、句法等信息。在构建规则中需要设计到大量的语言学知识，且可移植性很差。

传统的统计机器方法利用大规模的语料来学习标注模型。如张祝玉提出了基于特征模板的随机条件场方法：特征模板是指人工定义的二值特征函数，提取特征所处的位置便是一个上下文的窗口，不同的特征模板之间可以组合形成一个新的特征模板[9]。而且通过 CRF 算法可以利用之前已经得到的标注信息和隐马尔可夫链模型的 Viterbi 解码算法来获得最优的序列，同时根据已知所设定的模板来定义符合条件的特征取值为 1，不符合条件的取值为 0，在训练阶段根据 CRF 算法建立标签的转移，使得在推导阶段对每个测试的句中位置进行标注。

近年来随着硬件计算能力的提升和词嵌入(word embedding)的表示，神经网络模型在很多 NLP 任务上都得到了广泛的应用，对于常见的序列标注任务（如中文分词、词性标注等任务）都可以采用将 token 从 one-hot 映射到低维空间的嵌入表示，然后将获得的 embedding 序列输入到 RNN 中，利用神经网络算法来自动提取文本特征，再使用 softmax 算法来预测每个分词(token)所对应的标签，这种端对端的模型训练使得可以关注数据和模型本身从而不依赖于特征工程，但是也同时存在着对网络参数的依赖大，模型的可解释性差等问题。并且也存在着输出的是非法的预测标签，如在标签 B-LOC 中是不可能紧跟着 I-PER 的，这是因为 softmax 在对每个 token 打标签中是独立的分类而不能利用前面已经预测的信息了。因此在模型的最后一层输出上加入 CRF 就能够解决这一问题。

接下来将介绍 NER 中的主流神经网络结构算法：

Collobert 在 NN/CNN-CRF 模型中提出了窗口方法和句子方法这两种结构进行命名实体识别，其中窗口方法仅使用当前预测词的上下文窗口进行输入，使用传统的 NN 网络结构进行训练；而句子方法则是使用整个句子作为当前预测词的输入，并且加入了词的相对位置特征来区分句中的每个词，最后再使用卷积神经网络 CNN 进行训练[10]。同时作者也提出了两种不同的训练目标函数：一种是基于传统词级别的分类问题进行标签的分类从而预测结果,另一种则是基于句子级别的序列标注问题计算 CRF 算法中的标签转移矩阵的得分。在该实验中，使用句子级别的 CRF 训练结果明显比词级别的分类问题的表现效果要更好。

Lample[11]借鉴于上述的 CRF 的思路使用 RNN 结构并且结合 CRF 层训练 NER 任务[11]。其模型结构从下到上分别是 Embedding 层（包括词向量，字符向量等其他特征），双向 LSTM 层（将不同方向所得到的结果进行拼接）以及最后的 CRF 层（序列标注进行评分），同时实验结果表明双向 RNN（LSTM 或者 GRU）在命名实体识别上的表现效果明显好于传统的基于丰富特征的 CRF 模型。这项研究也充分利用了深度学习算法的优势，即无需特征工程仅利用字词向量就可以得到较好的效果，如果能引入高质量的词典特征（如 gazetters 地理

学词典）就能更进一步的提高训练结果。

最新的基于神经网络的 NER 研究主要通过引入注意力机制(Attention Mechanism)或者是对少量的标注数据的预训练来提升模型的训练效果。其中 Rei 就通过 attention 机制将原始的字符向量和词向量的简单拼接改进为权重求和，并且使用了两层神经网络来学习 attention 的权值，这样模型就可以动态的结合字词向量的信息了[12]。而 Dong 则在原始的 BiLSTM-CRF 模型上引入了汉字的字型结构（如“朝”被分解为“十日十月”），根据在 radical-level 的信息来进行序列标注，从而获得更多的信息和更深层次的细节[20]。 深度学习需要大量的训练数据，但对于 NER 而言数据标注集并不多，因此使用少量的标注数据也成为研究的重点。Matthew 就利用了半监督学习从海量无标注数据中训练一个双向神经网络模型，然后使用该模型获取标注词的语言模型嵌入(Language Model Embedding)，把这个向量作为特征加入到原始的 RNN-CRF 模型中可以大幅提升 NER 的训练效果[13]。此外还有 Yang 提出的迁移学习的方法也可以使用少量标注数据来研究 NER 领域[14]。

1.4 论文背景及意义

自改革开放的短短的几十年时间里，中国的快速发展和经济上的增速让世界瞩目，中国也在政治、经济、文化等方面发挥着越来越重要的影响力。在享受着这份身为华人的自豪与荣誉的同时，我们也能察觉到其他国家、民族、恐怖主义组织或者团体的或敌或友的态度，例如最近的硝烟逐渐浓烈的中美贸易战，就将很有可能因为直接冲突所唤起的社会情绪而引起摩擦甚至战争。置身于全球化浪潮中，我们需要响应习近平总书记提出的“总体国家安全观”的新政策，强调对“恐怖主义、分裂主义、极端主义”的零容忍，关注周边地区对国家安全产生影响的外部条件和内部条件[6]，以迅雷不及掩耳盗铃儿响叮当仁不让之势从体制和机制上遏制输入型的安全威胁和战争以及坚持使用铁的手腕予以沉重的打击。

如何从海量的数据中挖掘出最真实可信以及提取出关键信息成为建立情报系统中最重要的一环，本课题拟采用命名实体识别(Named Entity Recognition)技术获取涉恐文本中的人名(Person)、地名(Location)、组织机构名(Organization)，自动从新闻中抽取和生成基于关键词检索的涉恐文本的实体标注列表。同时分别对传统的基于规则的 CRF 算法和近些年来更流行的神经网络算法进行中文命名实体识别任务进行算法设计和评测来计算准确率率、召回率、F 值，从而对算法提出改进方案。

一方面，从语言分析的角度来看，命名实体识别任务属于中文分词中未登录词识别的基础问题，它是未登录词中识别难度最大的问题，也是在信息抽取、机器翻译等多种 NLP 任务中必不可少的组成部分，因此它的研究成果将直接影响到文本信息自动化处理的深层次研究[7]。另一方面，提高在涉恐的中文文本信息处理的能力和结合深度学习的前沿方

法是 NLP 领域的研究趋势，传统的基于规则的识别是通过领域专家和语言学者手工制定的规则，但这种方法仅适用于简单的识别系统并且要求规则之间不能发生冲突，因此制定规则会消耗大量的时间和精力且领域的迁移性欠佳。人们也越来越认识到在自然语言处理领域中，问题所输出的解不是相互独立的，而是在时间或结构上存在相互依存的结构化标注，包括序列、树状等图结构。使用序列模型算法就可以充分利用这一信息，并且结合神经网络的结构使用更少的手工构建依赖和更快的训练速度能够从而更加精确且高效的识别中文实体。

基于以上两个方面的考虑，本文的主要任务是获取涉恐文本，中文命名实体识别的常用框架和算法在已标注文本中的评测和分析，以及自动从涉恐文本中抽取中文命名实体。

1.5 本文工作及论文结构

本文的主要工作是通过利用关键词索引从百度搜索获得涉恐文本内容，对涉恐文本进行数据清洗及整理，然后基于 NER 的算法和框架，对已标注的中文命名实体识别语料进行评测和比较，并且最终自动化抽取涉恐文本的命名实体关键词组成文本检索列表。

本文一共分为五章。

第一章为引言部分，描述了涉恐文本和命名实体识别的定义和研究意义以及国内外研究现状，并概要说明了本论文的主要工作。

第二章先从全球恐怖主义数据库的介绍入手，阐明了本论文的数据来源、数据预处理以及期望生成的格式化内容。

第三章为中文命名实体识别框架的评测部分，分别就 Boson NLP 和 LTP 的两大框架进行评测和结果分析。

第四章为算法设计和评测部分，分别对基于模板的 CRF 算法和神经网络算法进行了深入的研究，并对算法评测的结果进行比较和分析，同时利用 LTP 框架实现了涉恐文本中命名实体的自动抽取。

第五章为本论文的最后的总结与展望部分，总结了本论文的主要工作内容及价值，并对之后的中文命名实体识别算法研究及涉恐数据库建设方向进行了展望。

2 涉恐文本的结构和搜集

2.1 全球恐怖主义数据库（GTD）概述

全球恐怖主义数据库（Global Terrorism Database）[1]是一个开源的数据库，其中收录了从1970-2016年超过16万件的全球恐怖主义事件信息，其建设经历了三个时期：

1） 1970-1997年的PGIS时期,PGIS负责对全球发生的恐怖袭击时间进行鉴别和记录，其记录的信息主要来源于通讯社的新闻报道、美国政府部门报告和外国报纸的新闻报道等。

2） 2001-2005年的GTD时期，PGIS将数据库交给马里兰大学电子信息化，马里兰大学的研究人员负责制定数据库编码，更新和扩展数据库并将这些数据录入到网络平台中。

3） 2006年至今的GTD2时期，该时期恐怖主义情报研究中心和应对全国联盟(START联盟)达成合作，确定了最终版的GTD2的收集指南[19]并更正数据收集的错误和弥补搜集过程中的质量缺陷，其中数据搜集小组被要求评估所搜集到的若干潜在事件，以此决定哪些事件是符合恐怖主义的标注，并通过来源文章的案例评估加以佐证，对于有歧义的案例将交给GTD2委员会作出最终决定。

为此，GTD 团队在定制文件管理和数据搜集工具的开发上建立了从数据搜集到编码入库的一整套流程，通过对来源文章的管理评估、案例的识别编码和人工的监督识别在不同层次上的严格的把控了信息的准确性，如图 2-1 GTD 数据采集流程图所示：

[1] GTD 数据库网址： http://www.start.umd.edu/gtd/

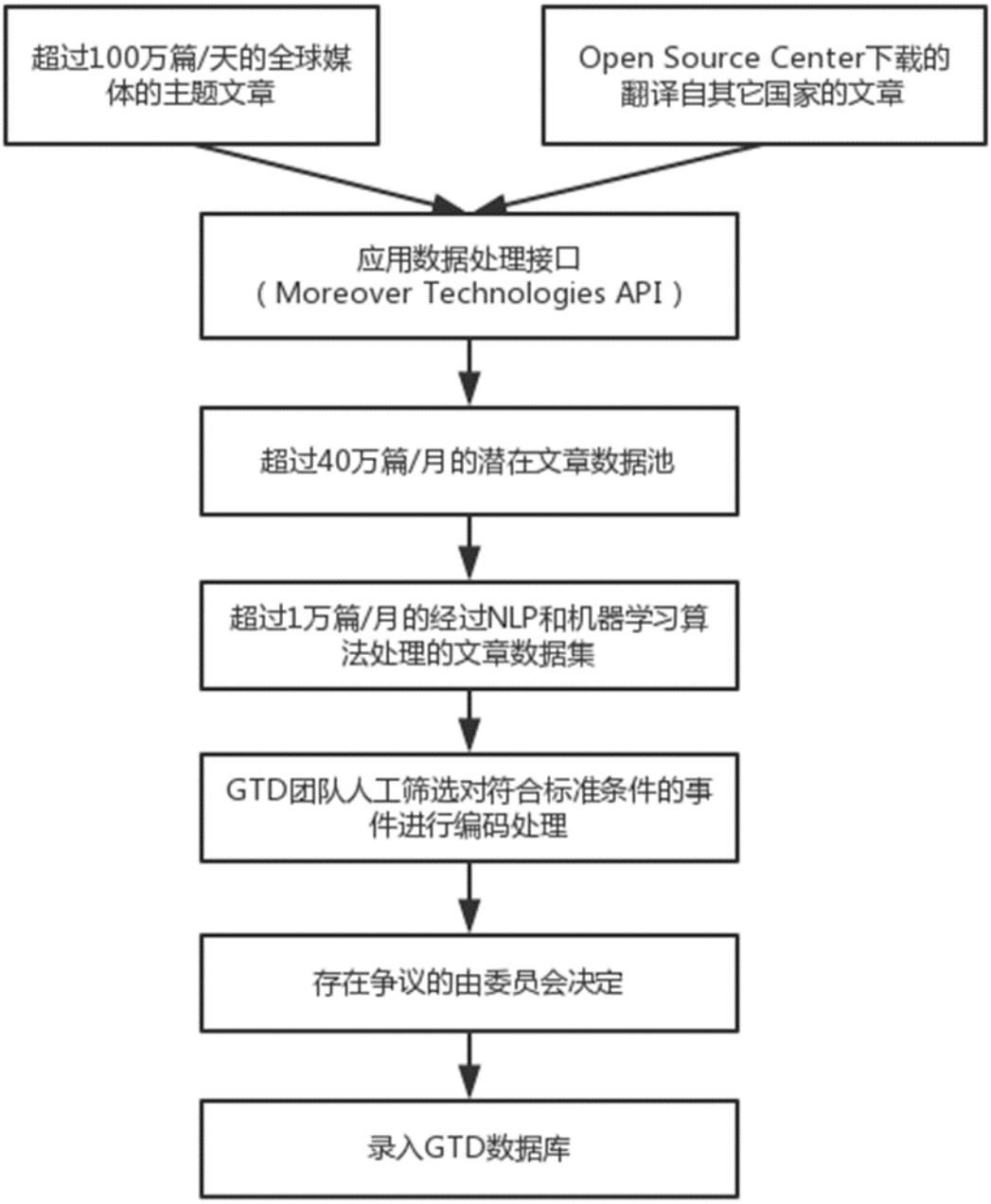

图 2-1 GTD 数据采集流程图[1]

虽然 GTD 能从时间和空间上显示事件在多维度上的发展和获取事件间更为准确的内在联系和丰富的知识，但仍存在以美为中心的政治意识偏见化。因此，建立专属的中文涉恐文本的情报分析成为我国国家战略安全部署的必要保证和刻不容缓的任务。

2.2 涉恐文本库的建立和清洗

2.2.1 涉恐文本的收集说明

本论文中的文本数据主要是从百度搜索引擎上基于关键字对新闻标题的内容进行爬虫，并将爬取的内容保存到 SQL 数据库中，文本的时间涵盖从 2000 年到 2018 年的新闻内容。

索引的关键词包含：武装组织, 袭击者, 劫持, 凶手, 绑架受害者, 人质, 恐怖组织, 恐袭, isis, 炸弹, 犯罪集团, 爆炸, 恐怖袭击等。

2.2.2 文本的格式说明及清洗

截止到 2018 年 3 月 14 日共爬虫所得的数据达 14 万余条，转换为 CSV 的表格形式后的每列数据存储格式为：

表 2-1 原始文本存储格式

列	title	date	keyword	source	content	link
说明	新闻标题	发布时间	关键词索引	报道来源	主题内容	爬虫链接

因为有的新闻主体内容为图片或者是视频，从而使得基于标题的爬虫所获得的数据为空，需要对这部分的数据进行删除且去重，最终余下 8638 篇文档。

2.3 涉恐文本的中文命名实体抽取格式说明

对于涉恐文本主要是获取其中的命名实体，包括人名、地名、机构名，而主体内容中则包含了很多其他的无关信息占用了大量的存储空间，因此使用算法仅获得命名实体并替代掉 content 中的内容，且另存储为单列内容，其格式如下：

表 2-2 实体抽取后文本存储格式

列	title	date	keyword	source	content	link	PER	LOC	ORG
说明	标题	时间	关键词	来源	内容	链接	人名	地名	结构名

在提取出对应的命名实体后，同时还保留了其他的关键信息，有利于追本溯源和统一文本内容。

2.4 本章小结

本章主要介绍了涉恐文本的数据来源、存储格式及特点，并通过预处理来过滤掉不符合规范的数据，同时还介绍了自动抽取所生成的目标格式。在下一章中，将详细介绍目前最流行的中文命名实体的两大应用框架并且将评测结果作为算法研究的基准 baseline。

3 命名实体识别的两大应用框架测评

3.1 训练集和测试集数据及测评方法说明

本论文的所有算法的测评都是基于同一个测试集[2]的，使用的是 BIO 标注体系。数据格式为每行中有以空格隔开的两列，分别是汉字（或字符），标注符号。

其中人名的标注符号：B-PER, I-PER 其中 B 代表开始，I 代表中间和结尾部分，合在一起之后就可以共同称之为人名，同样的还有地名：B-LOC, I-LOC 以及 组织机构名：B-ORG, I-ORG。

本论文的评测标准是基于 Sighan Bakeoff -3[3] 中的命名实体识别任务的标准，对测试集和算法标注的结果进行两两比较分别对人名、地名、组织机构名以及总体的准确率（Percision）、召回率（Recall）以及 F1 值（F1-Measure）进行计算。

$$准确率 = \frac{算法标注正确的实体总数}{算法标注的实体总数}$$

$$召回率 = \frac{算法标注正确的实体总数}{测试集中所有的实体总数}$$

$$准确率 = \frac{2 * 准确率 * 召回率}{准确率 + 召回率}$$

3.2 玻森中文语义平台（Boson NLP）的命名实体识别测评

3.2.1 Boson NLP 的使用说明

玻森中文语义 API[4]提供多种语义分析的 API，包括情感分析，新闻分类等方法，对每个登录注册的开发者提供了每天 500 次的命名实体标注的 API 使用机会，其中每次处理的文本不超过 5000 个字符。

---

[2] 测试集数据来源： https://github.com/crownpku/Small-Chinese-Corpus/tree/master/NER_chi

[3] 评测方法说明： http://sighan.cs.uchicago.edu/bakeoff2006/

[4] Boson NLP 网址： https://bosonnlp.com/

表 3-1 Boson NLP 识别实体类型

时间	地点	人名	组织名	公司名	产品名	职位
Time	Location	Person_name	Org_name	Company_name	Product_name	Job_title

则在实际训练中，并不进行细颗粒度的划分，因此将识别出 Company_name、Product_name 和 Org_name 合并为 ORG 命名实体，Person_name 为 PER 命名实体，Location 为 LOC 的命名实体。

表 3-2 BosonNLP.ner()的返回结果

返回值	类型	说明
(s, t, entity_type)	三元组的列表	word[s:t]的 entity_type 的实体
tag	列表	词性标注的结果
word	列表	分词后的结果

3.2.2 测评结果分析

表 3-3 Boson NLP 评测结果

	Precision 准确率	Recall 召回率	F1-Measure F1 值
PER 人名	90.68%	85.09%	87.79
LOC 地名	83.46%	60.69%	70.28
ORG 组织机构名	61.54%	72.00%	66.36
Total	77.22%	69.80%	73.32

根据表格可知对于准确率人名的表现效果最好，而组织机构名的表现最差，考虑到这是把上述的”Company_name”、”Product_name”计算进来所导致的误差，如“苹果手机”就被识别为产品名从而被认为是组织机构名而增加了识别误差，但是有些如“苹果公司”被识别为公司名，但同时也是组织机构名因而被识别正确。 而对于召回率，地名的表现最差，这是因为中文的地名往往很长而且并不具有规律性，难以用基于规则的方法来准确识别，如果引入地名词典的话则能够进一步的提高识别的准确率。

3.3 哈工大语言云（LTP）的命名实体评测

3.3.1 LTP 的使用说明

哈工大语言云技术平台[5]则提供了最基本的三种实体类型（PER, LOC, ORG）的识别，但也可以自主设计将实体类型拓展成品牌名、软件名等其他实体类型。

Pyltp 作为 LTP 的 Python 封装，提供了分词，词性标注，命名实体识别，依存语法分析等功能，在对文本进行命名实体识别时使用了如下的方法和模型：

文本分句： 使用 SentenceSplitter 进行分句

句中分词： 使用 Segmentor 调用 cws.model 分词模型进行分词

词性标注： 使用 Postagger 调用 pos.model 标注模型进行词性标注

命名实体识别： 使用 NameEntityRecognizer 调用 ner.model 模型进行实体识别

其中 LTP 采用 BIESO 的标注体系，并且输出的命名实体标记分别为 Nh（人名）、Ni（组织机构名）和 Ns（地名）然后再把上述的标记换成对应的 BIO 标注符号进行评测。

3.3.2 测评结果分析

表 3-4 LTP 评测结果

	Precision 准确率	**Recall 召回率**	**F1-Measure F1 值**
PER 人名	96.10%	89.97%	92.93
LOC 地名	85.30%	73.62%	79.03
ORG 组织机构名	82.21%	50.66%	62.69
Total	87.65%	71.05%	78.48

由表 3-4 可知，相对于玻森 API 的评测结果 LTP 的结果明显更优，但在组织机构名的召回率上的表现则更加糟糕，这是因为 LTP 在进行分词操作的时候把更长的组织机构名切分开来进行标注和实体识别。如“香港普乐斯公司”就被切分为“香港”“普乐斯公司”从而导致识别错误，因此中文分词的误差就会被累积从而对命名实体识别引起更大的误差，但如果采用 LTP 所支持的分词自定义词典就能够降低这一误差值。

[5]哈工大语言云平台 LTP 网址：https://www.ltp-cloud.com/

3.4 本章小结

本章主要介绍了玻森 NLP 和哈工大语言云两大中文命名实体识别框架，首先说明了测试集来源及格式、评测方法和框架的使用规则的设计，然后对于测试集进行数据上的评测分析，基于结果进一步探讨了提升结果的可能方法和策略。从中文的分句、分词、词性标注、命名实体识别的流程可以更容易的去理解算法运行的流程但同时也看到每一步的处理都将引入误差。在下一章中，将对 CRF 随机条件场算法和神经网络算法进行评测分析。

4 命名实体识别算法的设计及评测

4.1 实验环境

本论文中使用的是基于 GPU 的 Linux 环境，对于神经网络部分使用的是以 tensorfow 为后端的 keras 版本以及 python3.5 的两块 GTX 1080Ti 的 GPU 进行训练，配置如下：

表 4-1 实验环境配置

Python	3.5
Anaconda	Anaconda3-4.4.0-Linux-x86_64
Tensorflow	tensorflow_gpu-1.4.0-cp35
Keras	2.1.5
Cudn	8.0
GPU 型号	GeForce GTX 1080 Ti
操作系统	Centos 6.5
库函数	Jieba, python-crfsuite, scikit-learn 等

4.2 基于模板的条件随机场算法（CRF++）

4.2.1 算法原理

随机条件场算法 CRF (Conditional Random Field Algorithm)是指在给定分词 token 序列 X 下（即观测序列$o_1, o_2, o_3, \ldots, o_n$）来计算对应的标注序列 Y（即隐状态序列$h_1, h_2, h_3, \ldots, h_n$）的概率值。

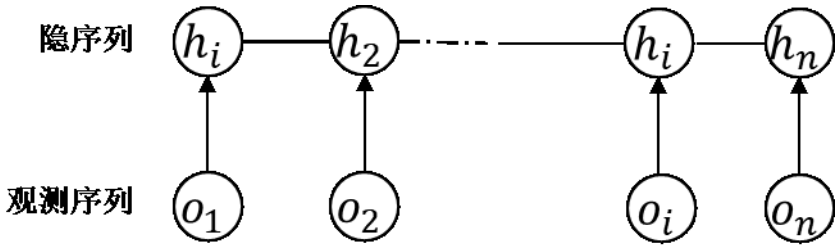

图 4-1 观测序列-隐序列

计算公式如下：

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l g_l(y_i, x, i)\right) \quad (4\text{-}1)$$

式 4-1 中的 $Z(x)$是规范化因子，λ_k和μ_l分别是针对特征函数f_k和g_l学到的权值，所得到的参数模型就是使得上述模型的概率最大化的参数。所预测的$\bar{y}$也就是使得满足$\max P(y|x)$条件的标注序列y。在预测算法中使用维特比(Viterbi)算法来求解概率最大路径[5]。

则定义 t 时刻到达状态 i 的最短路径为：

$$\delta_t(i) = \max P(i_t = i, i_{t-1}, .., i_1, o_t, ..., o_1|\lambda), i = 1,2, ..., N \quad (4\text{-}2)$$

式 4-2 中i_t表示最短路径，o_t表示观测序列，λ表示参数模型。

则可以定义递推式为：

$$\delta_{t+1}(i) = \max[\delta_t(j)a_{ji}]b_i(O_{t+1}), i = 1,2, ..., N; t = 1,2, ..., T-1 \quad (4\text{-}3)$$

式 4-3 表示 t 时刻处于状态 j, t+1 时刻转移到状态 i 且观测符号为O_{t+1}的最大概率。

时刻 t 到状态 i 的概率最大路径的经过的最短路的结点为式 4-4。

$$\psi_t(i) = \underset{1 \le j \le N}{\operatorname{argmax}}[\delta_{t-1}(j)a_{ji}] \quad (4\text{-}4)$$

最终返回的路径$y_i^* = \psi_{i+1}(y_{i+1}^*)$，全局最优路径则为$y^* = (y_1^*, y_2^*, ..., y_n^*)$

4.2.2 算法实现

采用条件随机场（CRF）进行命名实体识别，使用了结巴分词库（jieba）进行中文分词和词性标注从而构建了文本的词边界和词性特征，在训练中使用了 50 次的迭代进行中文分词和词性标注，然后对标注的文本句子进行命名实体识别并进行评测。使用字、词语边界和词性标注的特征模板来组合特征从而进行训练和评测。

“结巴”分词是一个 Python 版本的中文分词组件，除了支持中文分词，同时也支持使用 jieba.posseg 用于标注分词中每个词的词性，用法如下：

```
In [4]: import jieba.posseg as posseg
        words = posseg.cut('我是北京师范大学学生')
        for word, pos in words:
            print ('%s %s' %(word, pos))

我 r
是 v
北京师范大学 nt
学生 n
```

图 4-2 结巴分词的使用示例图

将标注后所得到的模型利用 CRF 将各个字所对应的实体标记分类结果进行命名实体识别，按照多分类任务对每个汉字进行实体的评价。其中 Python-crfsuite 是 CRFsuite 的 Python 版本，CRFsuite 是著名的条件随机场的开源工具，支持结巴词性标注后的输出作为特征的选取部分，并通过选取上下文的窗口来进行模板的组合，在实际训练中利用 Python-crfsuite 的 Trainer()进行模型的训练，Tagger()进行实体的标注。

4.3 基于字符的长短记忆网络算法(LSTM)

4.3.1 算法原理

LSTM(Long Short-Term Memory)是循环神经网络(RNN)的一种特殊形式，它作为一组前向网络的组合，每个单元代表一个时间节点，在每个单元(cell)内部分别有输入门、遗忘门和输出门，可以用于解决长序依赖问题。这里使用了双向的 LSTM 网络，即两层互为相反方向的输入，其最终的输出向量被理解为输入的网络表达方式，在最终的标注中一般使用 softmax 函数映射处理，但在实际的序列标注任务中，标签的上下文存在依赖关系，因此在输出端使用 CRF 可以有效的利用句子级别的标记信息[18]。

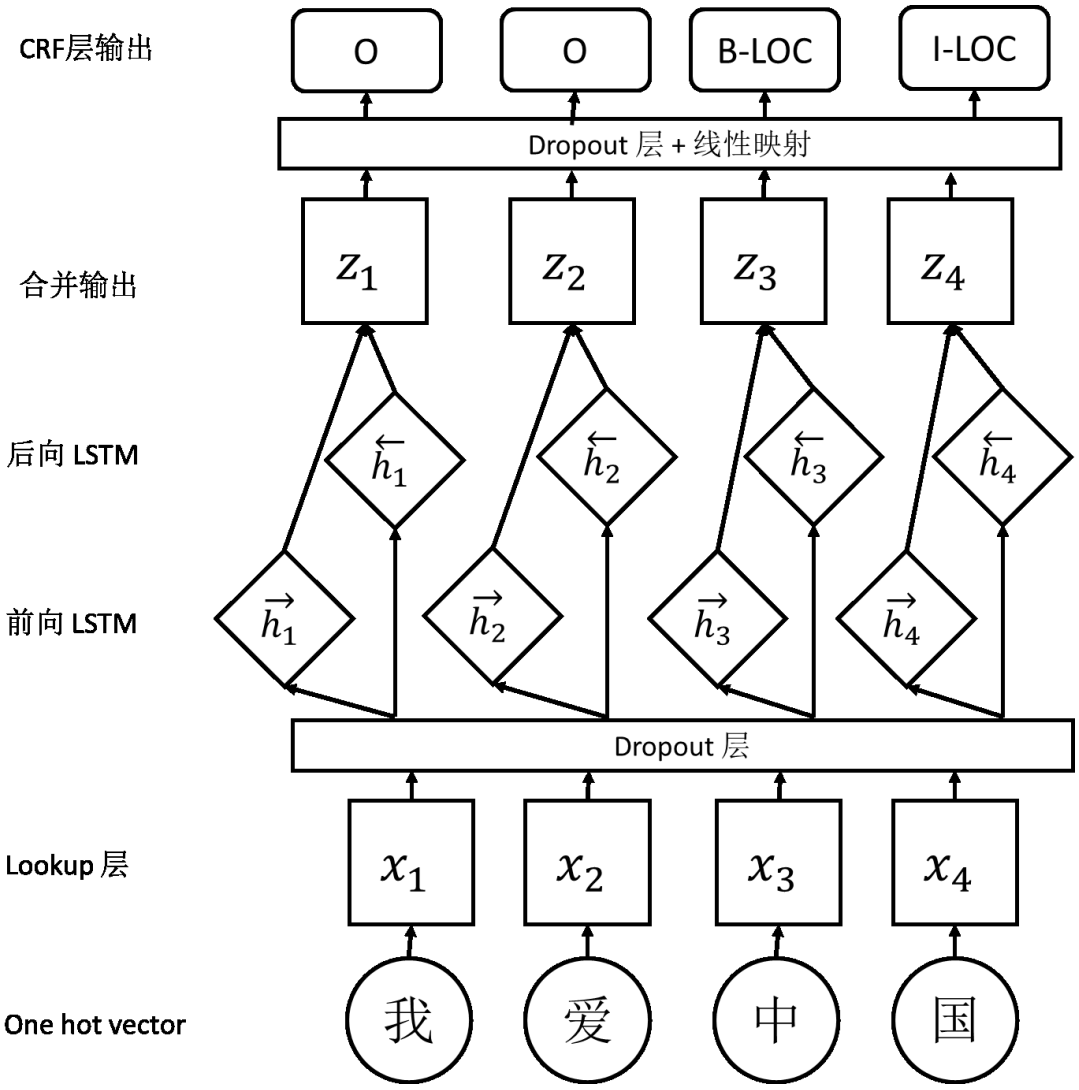

图 4-3 LSTM 模型构成示意图

对一个句子$s = (w_1, w_2, w_3, \dots, w_n)$中每个$w_i$表示句子的第 i 个字在字典中的 id 值，先离散化得到对应的 one-hot 向量。

在 look-up 层使用初始化的 embedding 矩阵将每个字w_i由 one-hot 向量映射为字向量 $x_i \epsilon \mathbb{R}^d$ (d 为矩阵的维度) 并且设置 dropout 来避免过拟合。

双向 LSTM 层用于自动提取文本中的句子特征，将句中各个字的$(x_1, x_2, x_3, \dots, x_n)$作为循环卷积网络层的各个时间步输入，分别得到正向输出$(\overrightarrow{h_1}, \overrightarrow{h_2}, \overrightarrow{h_3}, \dots, \overrightarrow{h_n})$和反向输出$(\overleftarrow{h_1}, \overleftarrow{h_2}, \overleftarrow{h_3}, \dots, \overleftarrow{h_n})$，并进行拼接得到$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$，(m 维)再设置一个 dropout 后接入一个线性层，将隐状态变量从 m 维映射到 u 维上（u 是标注集的标签数），获得提取到的句子特征

矩阵$\mathbf{Z} = (z_1, z_2, z_3, \dots, z_n)$（n*u 维），对于 Z 中的每一维$\mathbf{Z}_{ij}$都可以看做是$\mathbf{x}_i$分类到第 j 个标签的打分值。

CRF 层用于进行句子级别的序列标注，参数是一个$(u+2)*(u+2)$的标签转移矩阵 A，其中A_{ij}表示的是从第 i 到第 j 个标签的转移得分，维数加 2 是因为用于分别标记句子的首尾状态，假设句子长度的标签序列为$\mathbf{y} = (y_1, y_2, y_3, \dots, y_n)$,则模型对于句子 s 的标签等于 y 的打分如下：

$$\mathrm{score}(\mathrm{x},\mathrm{y}) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=1}^{n+1} A_{y_{i-1}y_i} \qquad (4\text{-}5)$$

由式 4-5 可知整个序列的分数等于各种位置的打分之和，而每个位置的打分分别是由 LSTM 的输出 P 和 CRF 的转移矩阵 A 来决定的，并使用维特比算法来求解最优路径。

4.3.2 算法实现

先使用预训练好的中文维基百科 embedding 来获得词典，再把训练集、验证集和测试集中的字符和标签分别用 200 维的词向量来表示并将使用数字来表示的字符和标签分别存为 x (字符),y(标签)。

然后构建序列模型，加入 embedding 矩阵，Dropout 层防止过拟合，使用双向的 GRU 来进行训练，因为 GRU 作为 LSTM 的变体，将输入门和遗忘门合并为一个更新门，同时也减少了更多的矩阵运算的计算量，再经过 TimeDistributed 层在每个时间步上操作了 Dropout 层来防止过拟合，最后经过 CRF 层实现命名实体的标注。

表 4-2 LSTM-CRF 层训练使用参数

Embedding 层输入维度	200	优化算法	Adam
GRU 层的输入维度	256	Batch_size 批训练大小	16
除 GRU 层间的 dropout	0.5	Epoch 训练次数	50
GRU 层间的 dropout	0.6	训练集维度	(21147 * 200)
CRF 输出层的维度	7	验证集维度	(2362 * 200)

4.4 基于字词特征的卷积和循环神经网络混合模型(CNN-LSTM)

4.4.1 算法原理

如上所述，使用基于字的循环神经网络（RNN）及其变形都能取得很好的结果，但如果仅依赖词向量则效果并不好，则可以使用卷积神经网络（CNN）来提取词的 n_gram 特征，本实验中使用 trigram 来提取 word embedding 矩阵的特征值，并且经过激活函数 LeakyReLU 线性修正后与 char embedding 的拼接作为输入[17]。

N-gram 模型是指若一个句子使用 s={w1,w2,w3…wn}表示，则计算句子 S 的先验概率：

$$p(s) = p(w_1) * P(w_2|w_1) * P(w_3|w_1w_2) * \ldots * P(w_m|w_1w_2 \ldots w_{m-1}) \quad (4\text{-}6)$$

但在实际计算中，使用马尔科夫假设来求得近似值，根据马尔科夫链对于随机状态的描述，某个状态只与其前的一两个状态有关系，更远的状态影响可以忽略不计，因此三阶的马尔科夫链（即三元语言模型）可将式 4-6 表示为：

$$P\{w_m|w_1, w_2, \ldots, w_{m-1}\} = P\{w_m|w_{m-1}, w_{m-2}\} \quad (4\text{-}7)$$

因此本实验中在上述的 charLSTM 中又增加了 wordLSTM，并将训练集作为模型的字序列输入，而将额外的 wordLSTM 作为辅助序列输入来进一步提升模型的准确率，并尝试对仅使用简单拼接和对于词嵌入使用 CNN 提取的 n_gram 特征融合的训练效果进行比较。

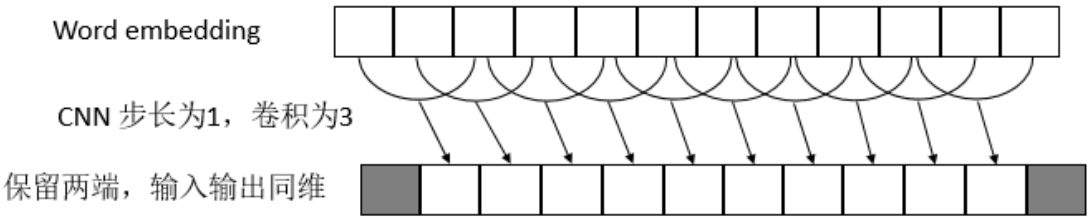

图 4-4 CNN 提取 n_gram 特征

得到经过卷积神经网络提取好的词级别特征后，将其与 char embedding 进行拼接。

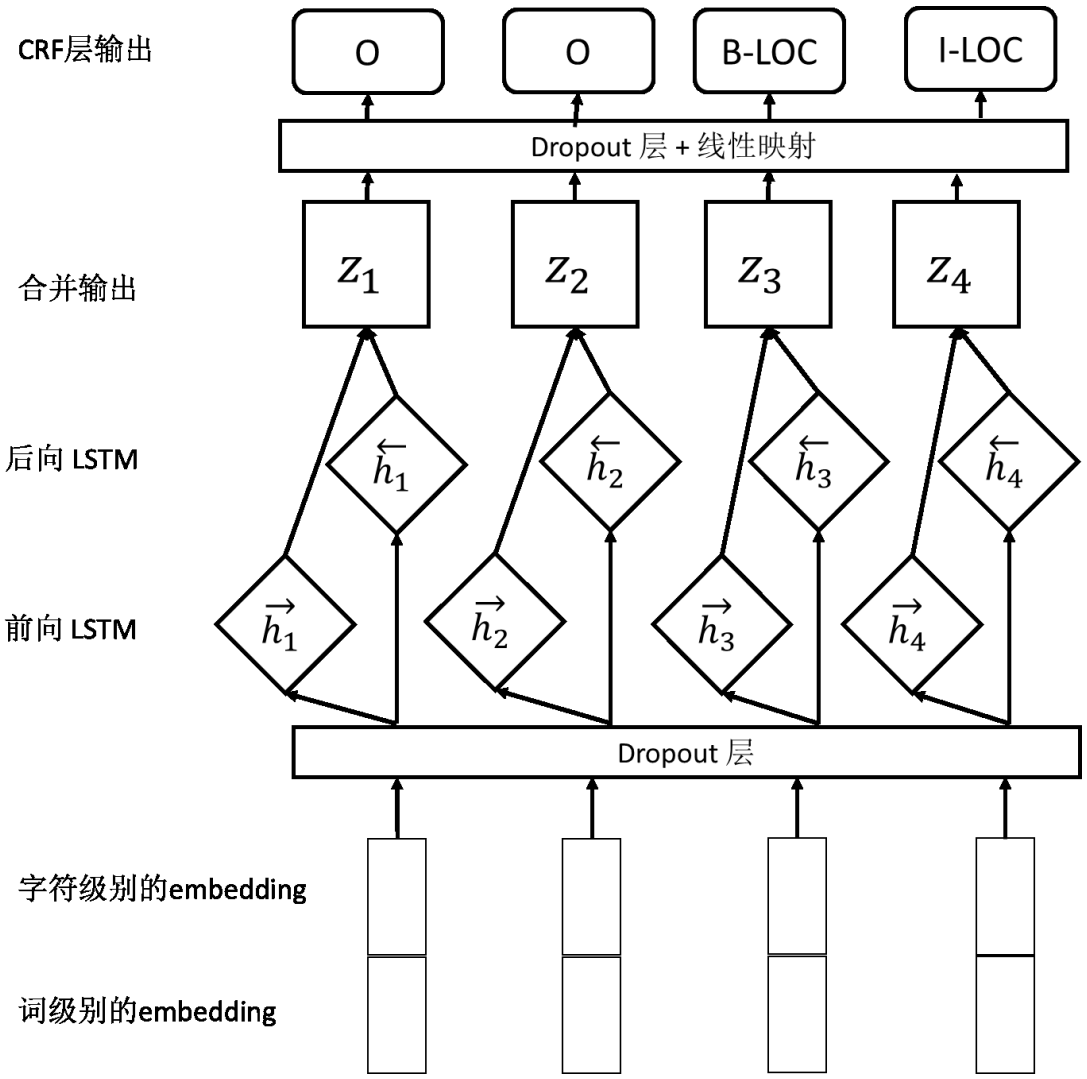

图 4-5 LSTM-CNN 模型构成示意图

4.4.2 算法实现

先分别得到训练集中的 word embedding 和 char embedding 部分，然后使用两种不同的方法进行计算，采用的第一种方式是利用 char embedding 和 word embedding 的直接拼接，再经过 CRF 层的多分类得到对应的标注集；第二种方式是使用卷积神经网络(CNN)来提取词的 n 元模型，其中步长为 1，卷积核长度为 3，通过 LeakyReLU 函数的线性修正之后再与 char embedding 拼接。

表 4-3 LSTM-CNN-CRF 层训练使用参数

Embedding 层输入维度	200	优化算法	Adam
GRU 层的输入维度	256	Batch_size 批训练大小	16
除 GRU 层间的 dropout	0.5	Epoch 训练次数	50

GRU 层间的 dropout	0.6	训练集维度	(21147 * 200)
CRF 输出层的维度	7	验证集维度	(2362 * 200)
卷积层的输出维度	128	卷积层核函数长度	3

4.5 评测算法的结果分析

表 4-4 所有的算法评测结果

框架	Precision 准确率	Recall 召回率	F1 值
Boson NLP	77.22%	69.80%	73.32
LTP	87.65%	71.05%	78.48
CRF++	83.69%	77.03%	80.22
LSTM-CRF	76.31%	74.76%	75.53
charLSTM-wordLSTM-CRF	80.22%	77.37%	78.77
charLSTM-wordLSTM-CNN-CRF	**88.94%**	**78.76%**	**83.54**

对于同一个测试集，分别运行表 4-4 所提及的命名实体识别的算法和框架，总的来说，字词级别的 LSTM 和 CNN 所提取的 n_gram 的综合模型无论是在准确率、召回率还是 F1 值都是表现效果最好的，但同时也发现经过仔细选择的基于特征模板的 CRF 模型的整体效果要好于 LSTM-CRF 的模型，而仅当加入了 n_gram 的窗口之后效果才更优了。则原因可能为：使用基于窗口的特征模板所提取到的特征要比神经网络的自动学习到的特征在长度较短且用字灵活的人名上更加的有效，而如地名和机构名的组成则长度较长且组成更加的复杂，因而双向的 LSTM 能够利用整句话的句子级别的语义特征而非仅特征模板所利用的简短窗口内的信息了。因此最后所设计的基于字词级别的 LSTM 和 CNN 就能够同时使用句子级别的语义信息和预测词的窗口信息从而表现的更好。

4.6 涉恐文本命名实体的自动抽取

在 2.3 章节中说明了涉恐文本的中文命名实体抽取的格式，结合上述的对于常用框架和算法的评测，选用哈工大语言云 LTP 的 API 接口来自动抽取文本中的实体，原因在于与其他的模型相比表现的效果较好，而且可以使用平台中已经训练好的模型更加快速的提取出结果出来，还可以跨平台(Windows, Linux, macOS)使用，而且训练消耗的资源很少，

而如 LSTM 或者 CRF++相关的模型虽然表现更好，但是均需要加载更多的内存资源，且在前期训练时就需要长达 2 天左右的时间，因此更加快速且准确的提取出命名实体就成为了最重要的任务。

提取出的结果图 4-6 所示，保留了基本的信息以及使用{PER，LOC，ORG}的词典形式保存在 content 内容中，并且将 PER, LOC, ORG 分列单独表示，从图中可知有些名词如“阿富汗”就被同时识别为地名和组织机构名，这是因为在不同的上下文语境中词语所表达的意思可能不同，则在不同的上下文窗口中所得到的结果也不同。对于所提取出的实体会发现地点往往跟新闻主题内容中发生的地点相联系。而组织机构和人名有时会跟新闻的报道机构和记者相关，因此在之后的设计中可以预先删除掉报道内容头部的信息避免产生干扰，同时可以引入常见的恐怖主义机构和恐怖主义犯罪分子的专有词典，进行进一步的细化处理，也可以建立关系图谱，对于含有同一个命名实体的部分产生链接关系以及来判断趋势走向。

	title	date	keyword	source	content	PER	LOC	ORG	link
33	喀布尔汽车炸弹	28/1/2018	袭击者	新浪新闻	PER: ['蒋超', '穆罕默德·穆斯塔	蒋超 穆罕默德·穆斯塔	喀布尔市 喀布尔 阿富汗	新华社 卫生部	http://news.sina.com.cn/o/2018-01-28/doc-ifyqyqni4086127.shtml
34	阿富汗首都发生	27/1/2018	袭击者	新浪	LOC: ['阿富汗']		阿富汗		http://video.sina.com.cn/p/news/o/doc/2018-01-27/200267903537.html
35	阿富汗首都喀布	27/1/2018	袭击者	凤凰网	PER: ['严王洁', '超峰']	超峰 严王洁	英国 喀布尔 中国 阿富汗	喀布尔 阿富汗	http://news.ifeng.com/a/20180127/55525813_0.shtml
36	荷兰首都枪击事	27/1/2018	袭击者	舜网	LOC: ['阿PER: ['李欢'] LOC: ['荷	李欢	荷兰 阿姆斯特丹 摩洛哥		http://news.e23.cn/guonei/2018-01-27/2018012700069.html
37	阿富汗东部城市	24/1/2018	袭击者	新浪新闻	PER: ['贾拉拉巴德', '贾拉拉巴德'	贾拉拉巴德 霍宇昂	南加哈省 阿富汗	路透社	http://news.sina.com.cn/w/zx/2018-01-24/doc-ifyqwiqk1119504.shtml
38	喀布尔酒店遭袭	22/1/2018	袭击者	中国新闻网	PER: [] LOC: ['京', '京']		京		http://www.chinanews.com/gj/shipin/2018/01-22/news752656.shtml
39	俄一周连发两起	19/1/2018	袭击者	澎湃新闻	PER: ['佩斯科夫', '普京', '	彼尔姆 普京 佩斯科夫	乌兰乌德市 莫斯科 俄 彼	中新社 俄新社 俄	http://www.thepaper.cn/newsDetail_forward_1959373
40	俄一学生袭击学	19/1/2018	袭击者	中华网	PER: [] LOC: ['俄		彼尔姆 俄罗斯 俄罗斯彼	俄新社 安全部	http://news.china.com/socialgd/10000169/20180119/31982797.html
41	俄罗斯一所学校	19/1/2018	袭击者	新浪新闻	PER: [] LOC: ['俄		俄罗斯 俄罗斯乌兰乌德	俄罗斯	http://news.sina.com.cn/w/2018-01-19/doc-ifyqtycx0133896.shtml
42	2名自杀式袭击者	16/1/2018	袭击者	大众网	PER: ['王安', '王	马安 王莉兰	瑞士 巴格达 伊拉克	法新社	http://www.dzwww.com/xinwen/guojixinwen/201801/t20180116_16923098.htm
43	伊拉克首都发生	15/1/2018	袭击者	东方网	PER: [] LOC: ['伊拉克', '		伊斯兰国 巴格达 伊 伊拉	新华社 伊	http://news.eastday.com/w/20180115/ulai11145256.html
44	伊朗一名袭击者	13/1/2018	袭击者	大众网	PER: ['小吴', '何 ', '何某	小吴 吴 何 何某	华西		http://www.dzwww.com/xinwen/guojixinwen/201801/t20180113_16910638.htm
	巴基斯坦遭自杀	10/1/2018	袭击者	新浪新闻	PER: ['阿巴西']	阿巴西	阿富汗 伊斯克省 北京	中新社 伊斯克省	http://news.sina.com.cn/o/2018-01-10/doc-ifyqeeia4552513.shtml

图 4-6 涉恐文本的中文命名实体抽取示意图

4.7 本章小结

本章首先对程序所使用的实验环境进行介绍，主要是基于 GPU 版本的 Linux 系统以及使用 tensorflow 为后端的 keras 深度学习框架进行快速的原型开发设计。然后就自主设计的条件随机场算法和神经网络算法进行训练和测试集的评测，并深入解释了每个算法的原理以及具体实现的过程。最后对评测集的结果进行分析以及实现了涉恐文本中的命名实体自动抽取，并提出了进一步完善反恐情报建设的可能性方案。

5 总结和展望

5.1 工作总结

随着网络信息的快速传播和发展以及严厉打击恐怖主义势力的国家安全战略，从网络的新闻媒体中获取恐怖组织信息以及活动时间和范围成为研究和建立反恐情报的重要途径。而针对恐怖主义袭击事件，最重要的是了解发生的时间、地点、任务和参与组织等，这些信息也是恰好可以被看作为命名实体进行信息的抽取。因此本文的主要工作就是先从海量数据资源中爬取涉恐文本的内容，对不符合规范的数据进行清洗和整理，并利用命名实体识别的算法和框架对标注的训练集训练和测试集进行评测比较，对评测结果进行理论上的分析。通过使用 LTP 的 API 接口实现快速准确的从涉恐文本中提取出命名实体并组合成文本索引的列表文件。

在具体的算法实现方面，使用了 Python 作为主要的开发语言，对 CRF 和神经网络模型进行研究分析，对于基于上下文窗口且进行模板组合的条件随机场算法在如人名的短文本中表现不错而对于名字更长的组织机构名则明显表现更差。而使用基于字符的双向 LSTM 算法则能够利用句子级别的语义信息但缺乏对于标注词的关注而使得学习到的特征容易受到干扰。因此最后所使用的 LSTM-CNN 算法则相当于同时结合了窗口特征和句子的语义特征，其中 CNN 算法是针对当前上下文的三个词进行特征上的提取，因此通过组合这两部分的信息能够进一步的捕获更多的特征从而提升命名实体识别的准确率和召回率。

通过以上的研究和分析，得知类似人名的更短且变换丰富的更适合使用窗口特征的方式进行特征的提取，而类似地名和组织机构名的更长且存在嵌套情况下的文本则更使用基于句子级别的语义信息进行特征的提取。同时对于中文的命名实体识别任务，如果能够使用自定义的中文分词词典就能够避免专有名词被切分开来从而不能被正确识别，提升识别的正确性。

5.2 工作展望

近几年，随着机器硬件性能的提升和深度学习的快速发展，神经网络算法在各大应用领域遍地开花，自从谷歌提出将文本用词嵌入的形式来表达，很多 NLP 领域的问题都可以使用深度学习的方法来解决了。而 NLP 领域中很多基础任务都可以看做成序列模型的标注问题，因为他们都可以使用概率图来表达变量相关关系，使用节点之间的边表示变量间的概率相关关系，并对于一个一维线性输入的序列来给序列中的每个元素打上标签集合中的某个标签。但其实很多时候都是没有现成的标注训练样本的，因此使用无监督学习或者半

监督学习就成为了研究序列任务的一个发展的方向和趋势。

全球化的背景下，建立涉恐领域的情报系统需要多个国家、组织的合作和资源共享。对于我国的情报系统而言，还需要公安有关部门就犯罪系统建立一整套完善的制度并且对于网络上的海量数据进行深度的挖掘从而扼杀恐怖主义活动于摇篮中。

参考文献

[1] 周松青. 全球恐怖主义数据库及对中国反恐数据库建设的启示[J]. 情报杂志, 2016, 35(9):6-11.

[2] 李本先, 梅建明, 李孟军. 我国反恐情报及预警系统框架设计[J]. 中国人民公安大学学报(社会科学版), 2012, 28(4):117-125.

[3] 魏静, 王菊韵, 于华. 基于多模块贝叶斯网络的恐怖袭击威胁评估[J]. 中国科学院大学学报, 2015, 32(2):264-272.

[4] 孙菲菲, 林平, 曹卓. 基于旋转森林集成学习的涉恐实体挖掘研究[J]. 情报杂志, 2015(5):190-195.

[5] 李航. 统计学习方法[M]. 清华大学出版社, 2012.

[6] 习近平就“亚洲安全观”作出哪些新论述?[EB /OL].[2016－05－15]. http: / /news.xinhuanet.com /world /2016－04 /29 /c_128944821.htm.

[7] 郑健. NLP 汉语自然语言处理原理与实践[M]. 北京：电子工业出版社, 2017.

[8] 宗成庆. 统计自然语言处理[M]. 清华大学出版社, 2013.

[9] 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]// 全国信息检索与内容安全学术会议. 2008.

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

[11] Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition[J]. Proceedings of NAACL-HLT. 2016: 260-270.

[12] Rei M, Crichton G K O, Pyysalo S. Attending to Characters in Neural Sequence Labeling Models[J]. arXiv preprint arXiv:1611.04361, 2016.

[13] Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, Russell Power. Semi-supervised sequence tagging with bidirectional language models[J]. ACL, 2017.

[14] Yang Z, Salakhutdinov R, Cohen W W. Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks[J]. ICLR, 2017.

[15] Lafferty J D, Mccallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]// Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 2001:282-289.

[16] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Readings in Speech Recognition, 1990, 77(2):267-296.

[17] Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[J]. 2016.

[18] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. Computer Science, 2015.

[19] START. GTD Codebook[R]: http://www.start.umd.edu/gtd/downloads/Codebook.pdf, 2017.

[20] Dong C, Zhang J, Zong C, et al. Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition[C]// International Conference on Computer Processing of Oriental Languages. Springer International Publishing, 2016:239-250.

致　谢

转眼间，本科的学生生涯即将结束，回顾四年所学所感所悟，有着诸多或欣喜或遗憾的时刻，然而当毕业论文即将提交的那一刻我也知道我的大学生涯也即将画上一个句号，在这四年的学习中收获颇多，感谢北京师范大学良好的学习氛围和各位老师们的严谨教学，我在这里学到了很多的做人、做事、做学问的道理，我也将秉承着这份踏实的学习态度继续前行。

首先我要感谢我的导师别荣芳教授，在本论文的撰写过程中，从论文选题、查阅资料、撰写开题报告再到设计实验以及最后的论文格式上的修改，别老师都提出了很宝贵的意见。当每次遇到难题时，我都会找别老师寻求帮助，而别老师不管或闲或忙都会积极的跟我面谈然后一起商量解决的办法。感谢这几个月来，别老师对我在学业上的精心指导以及对于我在未来发展上的关怀和帮助，无论是格式规范还是论文组织结构别老师都不厌其烦的指导我，使得我能够最后顺利的完成毕业论文，我所取得的点滴进步无不凝聚着别老师的心血。同时别老师国际化的视野、严谨勤奋的学术态度、杰出敏锐的思维都给我深深的启迪。

还要感谢母校——北京师范大学，都说巨人是站在肩膀上的，正是因为双一流的学术环境和深厚的历史文化底蕴，让我在学业上和人格品质上都得到了极大的成长。北师大作为国际知名学府也在不断的往国内外输送顶尖人才，感谢北师大的平台让我有机会分别在香港理工大学和加拿大英属哥伦比亚大学进行各为期三个月的学术交流合作，这些宝贵的机会不仅拓展了我的视野也让我学会谦卑学会向更加优秀的人学习。尤其衷心感谢学院的孙波老师、于小雷老师、孙云传老师和各位同门师兄弟姐妹们给我的帮助，正是你们的无私帮助让我度过一个又一个的难关，也正是学院所有老师们辛苦的付出和毫无保留的奉献才教育出了北师大一批又一批优秀的学子，也让我时常怀着一颗感恩的心去珍惜周围的一切。也暗自庆幸和四年的同学们度过的这段青葱岁月，给自己的人生添上一笔靓丽的彩色。

我还要特别感谢寒假实习的网易有道人工智能组的黄瑾老师和付凯学长，正是这段实习期间让我感受到了学术界和工业界在技术研究上的巨大差异，在这段期间我也认识到了自己在工程代码能力上的不足，也让迷茫的我渐渐的找到了方向，在一步步的试错中学习和收获成长。同时还要感谢在香港和加拿大所认识的同学、朋友和老师们，虽然感伤以后再见很难了，但这是这些过去的经历才塑造了现在的我，也让我学会珍惜共同度过的美好时光。

最后，我要感谢我的父母，一直在背后支持着我，让我学会感恩，学会自立，让我能够自由快乐的成长。我现在也认识到了一个人应当肩负的责任和使命，而我将带着这份“宁静致远”的初心坚持走下去，每一步，用心走，不放弃！