

**学术学位硕士研究生学位论文中期报告**

论文题目： 面向新疆暴恐事件的命名实体识别探究

|  |  |
| --- | --- |
| 姓 名： | 林广和 |
| 学 号： | 21509174 |
| 学科专业： | 计算机科学与技术 |
| 指导教师： | 张绍武 |
| 入学日期： | 2015年9月7日 |
| 报告日期： | 2017年07月？？日 |
| 报告地点： | （待定）创新园大厦A802 |

# 开题时拟定的研究方案、进度计划

## 1.1 开题时拟定的研究方案研究背景

本研究的研究目标为：《面向新疆暴恐事件的命名实体识别研究》。近年来，国际恐怖主义猖獗，导致世界各地伤亡惨重，震惊世界的 2015 年发生的法国巴黎恐怖袭击事件和 2016 初的比利时布鲁塞尔恐怖袭击事件尤其严重，两国社会和周边国家造成极大的冲击。同时，西方反华势力纵容支持“世维会”等境外“东突”势力，同时利用网络平台传播政治谣言对我进行攻击。在大数据时代下，新型的分析技术和相关工具，量化舆论倾向性，构建不同的模型，为舆论倾向性的研究提供新的思路。目前机器学习方法日益成熟，其应用的深度和广度都得到扩展，传统的统计学习和新兴的深度学习得到了广泛的应用，为解决命名实体识别等提供了较好的方法和模型。因此，研究基于暴恐事件的深度学习命名实体识别具有重要的现实意义。

本研究的主要内容有：通过爬虫技术采集新疆相关的新闻语料，对语料文本进行耗时耗力的人工标注，进行数据预处理，深度学习模型构建，最后进行命名实体识别的分类任务。

具体研究内容包括：

1. 暴恐事件的语料采集工作。通过谷歌全球数据库（GDELT）[2]获取相关链接，根据分析后采集与新疆相关的新闻文本。
2. 语料的标注工作。调研1999年的《实体识别定义》[1]一文，根据实体识别标注规范，对语料进行人工标注。
3. 词向量训练。利用英文维基百科公开的数据进行词向量的训练。由于维基百科的数据是 xml 格式，因此需要一系列的预处理，将 wiki 数据转换为 text 格式，然后使用 word2vec 的 python库gensim进行词向量的训练。
4. 引入字符向量。在常规的Bi-LSTM的基础上，引入基于CNN的字符向量，以学习更加细粒度的词特征。
5. 引入CRF层。在模型输出接入 CRF 层，进行模型训练，以确保全局最优。
6. 利用事件抽取技术，以结构化的形式抽取和存储新疆暴恐事件。

## 进度计划

针对以上研究内容，原定的进度计划为：

2016年9月至2016年10月，广泛阅读文献，了解国内外研究进展，查阅国内外新方法，新思路，调研课题并提高方案可行性。了解基本的深度学习模型，学习命名实体识别的相关概念。

2016年11月，初步拟定研究方案，并论证方案可行性。语料采集，并根据语料的实际情况进行相关的数据清洗工作。

2016年12月，搭建深度学习框架Keras的工作环境，进行语料标注。

2017年1月，实现字符向量模块，实现句子级别对数似然函数

2017年3月，进行实验，训练和测试，优化NER模型。

2017年4月至2017年10月，完成事件抽取。

2017年11月至2018年3月，进行系统测试

2018年4月，完成论文初稿。

2018年6月，完成论文终稿。

# 学位论文的研究进展完成情况、阶段性成果和创新点论述

## 2.1 研究进展完成情况

根据开题报告确定的研究内容、研究方案、技术路线以及进度安排，现在已完成的内容如下所述：

1. 完成数据采集，首先从GDELT网站下载并导入2013年9月以后的新闻数据库， 根据链接分析，获取有用站点，在有用站点基础上，定点爬取数据1000余篇以供标注。
2. 根据文献[1]中的标注规范，独立完成对语料的标注和校对工作。
3. 利用英文维基百科公开的数据完成词向量的训练。将维基百科的数据进行预处理后使用python版的gensim进行词向量训练。
4. 使用深度学习框架Keras构建Bi-LSTM+CNN+CRF模型，复现文献[4]实验。
5. 利用事件抽取技术，以结构化的形式抽取和存储新疆暴恐事件。

从以上论文覆盖的研究内容来讲，完成(1)-(3)中的工作，由于(4)中的程序出现Loss值无法下降的问题，正在调试当中，故研究进度暂时停留在调试程序中，导致(5)的工作尚未展开。

## 阶段性成果展示

### 数据采集

GDELT每个15分钟提供全球事件数据。GDELT目前的事件库约有3.5亿条事件数据。这些事件从1979年1月1日开始一直到今日。GDELT第一项服务就是免费的数据下载。同时GDELT还在谷歌的BigQuery上提供了数据API，这样可以使用谷歌的分析工具进行分析。GDELT的数据除了事件数据外，还提供了GKG数据，也就是全球知识图(Global Knowledge Graph)的数据。

通过GDELT提供的相关链接，可下载zip类型的GDELT数据压缩包，解压后为csv格式，每个csv文件内的数据均有58个字段。通过导入数据库，进行关键词查询，筛选出与本研究相关的新闻，保存相关新闻链接，以便于进一步的采集。

网络爬虫（Web crawler），是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本，它们被广泛用于互联网搜索引擎或其他类似网站，可以自动采集所有其能够访问到的页面内容，以获取或更新这些网站的内容和检索方式。从功能上来讲，爬虫一般分为数据采集，处理，储存三个部分。

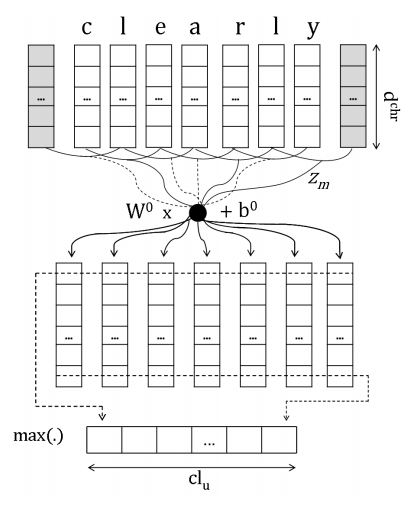
本研究分析GDELT上的链接，决定采用定点爬虫的方式采集新闻语料。使用WebMagic这一无需复杂配置、便于二次开发的爬虫框架，它提供简单灵活的API，只需要少量代码即可实现一个爬虫。该框架采用完全模块化的设计，功能覆盖整个爬虫的生命周期（链接提取、页面下载、内容抽取、持久化），支持多线程抓取、分布式抓取，并支持自动充实、自定义UA/Cookie等功能，通过一些简单的设置，避免了一些网站反爬虫的限制。

### 语料标注

根据文献[1]提供的标注规范，结合通用数据集CoNLL 2003的标注方式，我采用B(Begin,实体的开始)、I（Internal，实体的中间部分）、E（End，实体的结束）、S（Single，代表该单词本身就是一个实体）、O（Other，其他）五个标注符号对语料进行标注。为了更好的区分人名、地名、机构名，我们定义了13种标记，L={B\_PERSON, I\_PERSON, E\_PERSON, S\_PERSON,B\_LOCATION, I\_LOCATION, E\_LOCATION, S\_LOCATION, B\_ORG, I\_ORG, E\_ORG, S\_ORG, O}，分别表示人名的开始、中间、结束、单独的人名、地名的开始、中间、结束、单独的地名、机构名的开始、中间、结束、单独的机构名、其他。IOBES这种标注体系有着很强的边界区分，便于模型的学习。

### 基于CNN的字符级词向量（CharWNN）

在命名实体识别中，每个实体在其形态学上均有特点，我们通过训练字符向量，找到其形态学的特征，具体方法如下：



给定一个词w，有M个字符，即 ，我们将每个字符 转换为对应字符向量 ，其中 ，则有 ，其中 采用的是one-hot编码。

卷积层的输入是字符的向量，由字符序列构成，形如 ，采用大小为 的卷积核进行级联后的卷积。具体计算如下：

设级联后的向量为

卷积层计算

其中 表示经过计算后的字符向量， 为卷积层权重，该权重用于抽取给定词窗口下的局部特征，再经过最大池化（MaxPooling）抽取该词的全局特征，这样就可以作为整个模型输入的一部分。

### 句子级别似然函数

在命名实体识别中，我们不能忽视标记之间存在依赖关系，而通常的单词级别的预测忽视了这种标签信息。我们采用Collobert等人[5]提出的句子级别似然函数方式：给定所有词对应的所有可能标签的预测和以及标签之间的转移得分，在训练时，我们尽可能降低错误路径的得分，提高正确路径的得分。

假设我们所训练的网络输出的得分矩阵为 ，矩阵元素 表示为网络在第t个词的第i个标签的得分。同时，我们引入转移得分矩阵A，用来表示从第i个标签转移到第j个标签的转移得分，同时转移矩阵做为网络的训练参数。整个句子关于标签的得分由转移得分和网络的输出得分两部分构成：

 。

我们对整个预测标签，即对所有可能的路径 进行归一化，则正确路径的条件概率为

,对该公式取对数，得到：



为了使公式（2）以线性时间计算，采用一下方法：

终止条件为



，最终我们能取得整个句子的最大似然。

而在推断过程中，我们希望最大化整个句子的得分，即



通过维特比算法，递归执行公式（1）、（2），替换公式中的logadd为max，同时记录每一次的最优路径便于回溯。

## 2.3 创新点论述

目前，由于上述开题时的想法已经有研究者[6]在ACL 2016上发表，所以经过大量调研和反复思考，重新提出未来科研的创新点及想法主要有：

1. 考虑到自己标注的语料规模比较小，而迁移学习通过已有的知识来学习未知知识，可以缓解语料不足造成的学习能力下降的问题。
2. 当下Attention机制广泛地应用于自然语言处理，尝试融入字符级Attention，从而提高实体识别的F值。

# 后续工作的设想、可能遇到的困难和问题及条件保障措施

## 3.1. 后续工作的设想

(1).语料的扩充

所采集的语料实体稀疏，部分文档中通篇只有一个相关实体，因而寻找合适的语料来源、扩种语料，并进行标注是十分必要的。语料的扩充和语料质量的提高有助于实验结果提升。后续会继续分析GDELT上的其他链接，完成爬取和标注工作。

(2). 创新点的实现

对创新点处提出的想法进行实验验证，研究字符级Attention机制和迁移学习是否能有效提高NER系统的实验效果。

## 3.2. 可能遇到的困难和问题

由于语料数量不足可能会导致实验结果不佳，但语料标注工作费时、耗力，而人力有限，无法做到标注和实验兼顾，需要自己调配好时间。

## 3.3. 条件保障措施

学校图书馆购置了充足的文献资源和丰富的图书馆藏，方便查阅学习，导师的回归为整个科研学习过程提供了新机遇。

# 已发表、录用的论文和已投稿的论文情况

# 参考文献（不占字数）

[1] Chinchor N, Brown E, Ferro L, et al. 1999 named entity recognition task definition[J]. Mitre & Saic, 1999.

[2] Leetaru K, Schrodt P A. GDELT: Global data on events, location, and tone[C]//ISA Annual Convention. 2013.

[3] Chollet F. Keras[J]. 2015.

[4] Ma X, Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf[J]. arXiv preprint arXiv:1603.01354, 2016.

[5] 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.

[6] Ma X, Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf[J]. arXiv preprint arXiv:1603.01354, 2016.