大连理工大学

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 副教授

入学日期：2015 年 9 月 7 日

报告日期：2017 年 7 月　日

报告地点：大连理工大学创新园大厦

研究生院制表

说　明

硕士学位论文中期检查是保证学位论文质量、工作进度和研究生培养质量的重要措施。原则上，要求硕士生应在第 4 学期末进行中期检查，其中，2 年制专业学位硕士生应在第 3 学期末进行中期检查。

一、考核内容：学位论文内容完成情况、阶段性成果是否正确，开题时方案是否需调整或已做了哪些调整，后续工作思路是否正确、工作进度是否有保障、预期目标能否实现、论文质量是否能够保证以及论文工作存在的问题等。

二、考核时间：原则上，硕士生的中期检查应在第 4 学期末进行，其中两年制专业学位硕士生的中期检查可在第 3 学期末进行。

三、报告撰写：中期报告正文字数不少于 4000 字，正文及参考文献等撰写要求参见《大连理工大学博士学位论文格式规范》。

四、考核办法：由学部（学院）组织，3-5 名本学科领域硕导及以上专家组成评审专家组以答辩的方式进行。学生进行口头陈述时间不得少于 10 分钟。专家组给出考核成绩和是否通过的意见。

五、报告保存：中期报告一式两份，签字后分别由学部（学院）和学生保存。

六、信息登录：中期考核后，学生应及时登录研究生信息管理系统上传中期报告（PDF 文档）及考核结果。

硕士学位论文中期检查报告正文

撰写大纲：

1. 开题时拟定的研究方案、进度计划；若开题时的研究方案已经调整，应说明调整的原因、调整后该领域的国内外研究状况分析、研究内容、研究方法、进度计划等；
2. 学位论文的研究进展完成情况、阶段性成果和创新点论述；
3. 后续工作的设想、可能遇到的困难和问题及条件保障措施；
4. 已发表、录用的论文和已投稿的论文情况。
5. 参考文献（不占字数）。

1 绪论

1.1 研究背景

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年初的比利时布鲁塞尔恐怖袭击事件尤其严重，给两国社会和周边国家造成极大的冲击。

在信息时代，以美国为首的西方反华势力纵容支持“世维会”等境外“东突”势力加大反宣渗透;境外敌对势力逐渐把网络阵地作为对我渗透攻击的主渠道;一些网络舆论突发事件背景复杂，部分突发事件被别有用心分子最终引向对党和政府的攻击、抹黑，对主流意识形态冲击较大。与此同时，西方还大力资助各种反华势力利用网络平台传播政治谣言对我国进行攻击。

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提供了较好的方法和模型。

因此利用机器学习和自然语言处理技术快速实时地分析新疆暴恐事件，及时了解其它国家的立场和情感倾向，对那些非客观、非公正、肆意歪曲的报道进行及时揭露，对维护国家形象以及打击极端恐怖分子的嚣张气焰起着至关重要的作用,对新疆暴恐事件的国际舆论倾向性分析具有重大的理论意义和现实意义。

利用计算机对新疆暴恐事件国际舆论倾向性分析应该遵循新闻事件分析的一般准则，首先进行新闻事件抽取即抽取出新闻六要素5W1H（Who, Where, What, When, Why, How）。计算机为了理解新闻各要素首先需要命名实体识别技术抽取出人名、地名、机构名等基本信息。在新闻事件抽取之后再进行进一步定量分析。

Tjong Kim Sang[10]、Tjong Kim Sang 和De Meulder[11]、Doddington[12]等都采用大量特征工程和其他NLP任务的结果进行实验，取得了先进的效果。

Ratinov and Roth[13]使用全局特征、来自维基百科的地名词典和类似布朗聚类式的词向量，在CoNLL-2003公开数据集上获得了90.80的F1值。

Lin and Wu[14]在不使用地名词典情况下，通过将搜索引擎查询记录库进行K-means聚类，提取短语特征用于NER任务，在性能上超过了Ratinov和Roth。

Passos[15]等人在只使用公开数据训练短语向量的情况下获得了近似的性能。

Suzuki[16]等人为了解决稀疏特征，采用大规模未标注数据进行降维，并在没有任何外部知识的情况下，构造了最先进的 NER 系统，其在 CoNLL-2003 上的 F1 值为 91.02。

Collobert[1]等人采用了深度神经网络模型进行联合学习，该方法采用 embedding 层和多层一维卷积的结构，用于词性标注（POS tagging），组块分析（Chunking），命名实体识别（Named Entity Recognition），语义角色标注（Semantic Role Labeling）等 4 个经典问题。[2] 文献[1]在 NER 训练时采用了句级对数似然函数，充分利用了标签之间的依赖关系，并获得了不错的效果。

Santos 等人[3]提出了 CharWNN 的网络，该网络是对 Collobert 等人提出的 FFNN 的一个补充，该模型在西班牙和葡萄牙语的 NER 中取得不错的效果。Labeau 等人[4]采用了带有字符级 CNN 的 BRNN 进行关于德语的序列标注任务。

Ma, Xuezhe[8]将 CharWNN 应用于英文的序列标注任务，同样取得了不错的效果。

由于近期的 NER 研究大量围绕于深度学习展开，而且非监督学习的自动学习特征有效避免了耗时费力的特征工程，所以决定使用基于深度学习的命名实体识别。

1.2 研究方法

在机器学习领域中，绝大多数先进的 NER 系统都采用需要大量人力的特征工程，以及依赖一些其他 NLP 的工具；而在采用深度学习的命名实体识别系统中，大多数采用了词向量作为模型的输入，以此减少像传统方法带来的维度灾难，同时最小化对特征工程的依赖。同时，文献[5]在词性标注上使用字符向量，对词进行形态学上的特征提取，命名实体识别和词性标注同为序列标注，且均是自然语言处理的一部分，因此，字符向量同样适用于命名实体识别。综上所述，我的原研究内容包括以下几方面：一、暴恐事件的语料采集工作。主要通过爬虫技术获取相当规模的暴恐事件语料，并对语料做适当的预处理。二、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[6]进行词向量的训练。三、采用深度学习框架 Keras[7]进行建模，在原始模型的基础上，引入字符向量。四、在模型输出接入 CRF 层，进行模型训练，

以确保全局最优。

1.3 进度计划

由于对课题难度估计不够，认识不深，开题报告中所制定的进度计划存在着很多不足之处，同时在阅读文献过程中，发现文献[8]已经实现开题时的想法，同时，在数据采集及标注过程中，发现相关实体稀疏的问题，因此我对进度计划时间、内容安排进行了修改和完善。目前，实验工作正按计划进行。其中 1-5 项已经完成，6-9 项正在进行中，具体内容如下：

1.2016 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。

2.2016 年 11 月，初步拟定研究方案，并论证方案可行性。

3.2016 年 12 月，搭建爬虫框架，语料采集，并根据语料的实际情况进行相关的数据清洗。

4.2016 年 12 月至 2017 年 1 月，2017 年 3 月，语料标注工作。

5.2017 年 4 月，调研 CRF 层原理。

6.2017 年 5 月至 2017 年 7 月，搭建深度学习框架 Keras 的工作环境，实现字符向量模块、句级对数似然函数，复现文献[8]实验。

7.2017 年 8 月至 2018 年 3 月,结合自己标注的语料,配合之后的调研工作，将新想法进行新的实验，训练和测试，优化 NER 模型。

8.2018 年 4 月，完成论文初稿。

9.2018 年 6 月，完成论文终稿。

2 研究内容概述

2.1 研究进展与阶段性成果

2.1.1 数据采集

GDELT 每隔 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。根据实际，从 GDELT 需要采集了 1100 余相关新闻语料用于标注。

2.1.2 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。目前已经完成采集得到的 1100 余篇语料的标注工作，准备用于实验。

2.1.3 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：

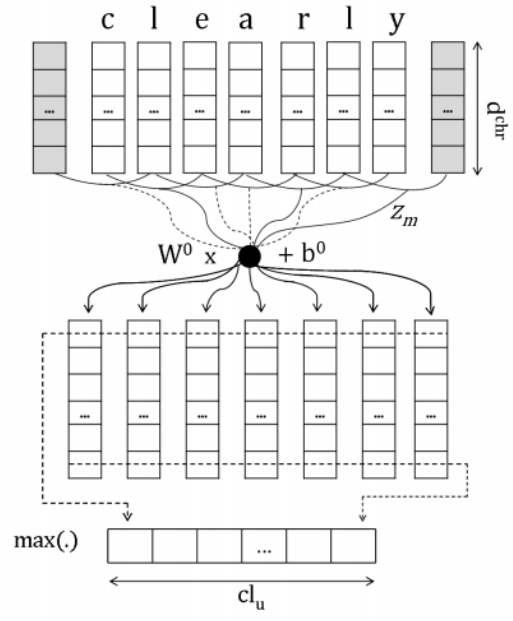

图 1.卷积方式获得字符向量

给定一个词 w，有 M 个字符，即 $\mathrm{w}=\{c_1,c_2,...,c_M\}$ ，我们将每个字符 c_m 转换为对应字符向量 $\mathrm{r}_m^{chr}\in W^{chr}$ ，其中 $W^{chr}\in\mathbb{R}^{\mathrm{d}^{chr}\times|V^{chr}|}$ ，则有 $\mathrm{r}^{chr}=W^{chr}v^c$ ，其中 v^c 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 $\{\mathrm{r}_1^{chr},\mathrm{r}_2^{chr},...,\mathrm{r}_M^{chr}\}$，采用大小为 k^{chr} 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 $z_m=(r_{m-(k^{chr}-1)/2}^{chr},...,r_{m+(k^{chr}-1)/2}^{chr})^T$

卷积层计算 $[r^{wch}]_j=\max_{1<m<M}[W^0z_m+b^0]_j$

其中 r^{wch} 表示经过计算后的字符向量，$W^0\in\mathbb{R}^{cl_u\times d^{chr}\times k^{chr}}$ 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。

2.1.4 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 $f_\theta([x]_1^T)$ ，矩阵元素 $[f_\theta]_{i,t}$ 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 $[A]_{i,j}$ 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 x_1^T 关于标签 i_1^T 的得分由转移得分和网络的输出得分两部分构成，即

$$s([x]_1^T,[i]_1^T,\tilde{\theta})=\sum_{t=1}^{T}([A]_{[i]_{t-1},[i]_t}+[f_{\theta[i]_t,t}])$$

我们对整个预测标签，即对所有可能的路径 $[j]_1^T$ 进行归一化，则正确路径的条件概率为

$$p([y]_1^T,[x]_1^T,\tilde{\theta})=\frac{e^{s([x]_1^T,[y]_1^T,\tilde{\theta})}}{\sum_{\forall[j]_1^T}e^{s([x]_1^T,[j]_1^T,\tilde{\theta})}} \quad (1)$$

对该公式取对数，得到：

$$\log p([y]_1^T,[x]_1^T,\tilde{\theta})=s([x]_1^T,[y]_1^T,\tilde{\theta})-\log\underset{\forall[j]_1^T}{adds}([x]_1^T,[y]_1^T,\tilde{\theta})$$

为了使公式（2）以线性时间计算，采用一下方法：

$$\begin{aligned}
&\delta_t(\mathrm{k})\triangleq\underset{\{[j]_1^t\cap[j]_t=k\}}{\mathrm{logadd}}\ \mathrm{s}([x]_1^t,[j]_1^t,\tilde{\theta})\\
&=\log\underset{i}{add}\ \log\underset{[j]_1^t\cap[j]_t=i\cap[j]_t=k\}}{add}\ \mathrm{s}([x]_1^t,[j]_1^{t-1},\tilde{\theta})+[A]_{[j]_{t-1},k}+[f_\theta]_{k,t}\\
&=\log\underset{i}{add}\ \delta_{t-1}(i)+[A]_{i,k}+[f_\theta]_{k,t}\\
&=[f_\theta]_{k,t}+\log\underset{i}{add}(\delta_{t-1}(i)+[A]_{i,k}),\forall k
\end{aligned} \quad (2)$$

终止条件为

$$\log\underset{\forall[j]_1^T}{add}\ \mathrm{s}([x]_1^T,[j]_1^T,\tilde{\theta})=\log\underset{i}{add}\ \delta_T(i)$$

，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

$$\arg\max_{[j]_1^T} s([x]_1^T,[j]_1^T,\tilde{\theta})$$

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。

2.2 创新点论述

目前，由于上述开题时的想法已经有研究者[8]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：

考虑到自己标注的语料规模比较小，受到文献[9]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。

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。

3 后续工作展望

3.1 后续工作的设想

（1）数据集的扩充

所采集的语料实体稀疏，部分文档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，以提升语料丰富性。

（2）实验方法的调研

现阶段仅仅从词的形态学信息上考虑问题，而并未考虑新疆实体在英文表达上呈现出其他的语言学上的特性。下一步将进一步调研相关内容，使得模型有更好的表达能力，以达到更好的效果。

3.2 可能遇到的困难与问题

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

学校图书馆及电子图书馆可以提供丰富的图书及文献资源，查阅便利；指导老师的回归在后续研究过程中将会提供悉心的指导与研究方向的精确掌控。

4　已发表、录用的论文和已投稿的论文情况

参考文献

[1] Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research, 2003, 3:1137-1155.

[2] 余凯, 贾磊, 陈雨强, 等. 深度学习的昨天, 今天和明天[J]. 计算机研究与发展, 2013, 50(9): 1799-1804.

[3] dos Santos C, Guimaraes V, Niterói R J, et al. Boosting named entity recognition with neural character embeddings[C]//Proceedings of NEWS 2015 The Fifth Named Entities Workshop. 2015: 25.

[4] Labeau M, Löser K, Allauzen A, et al. Non-lexical neural architecture for fine-grained pos tagging[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 232-237.

[5] dos Santos C N, Zadrozny B. Learning Character-level Representations for Part-of-Speech Tagging[C]//ICML. 2014: 1818-1826.

[6] Rehurek R, Sojka P. Software framework for topic modelling with large corpora[C]//In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. 2010.

[7] Chollet F. Keras[J]. GitHub repository: https://github.com/fchollet/keras, 2015

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

[9] Yang Z, Salakhutdinov R, Cohen W W. Transfer learning for sequence tagging with hierarchical recurrent networks[J]. arXiv preprint arXiv:1703.06345, 2017.

[10] Sang EFTK. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition[J]. Computer Science, 2002:142--147.

[11] Tjong Kim Sang E F, De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition[C]//Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, 2003: 142-147.

[12] Doddington G R, Mitchell A, Przybocki M A, et al. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation[C]//LREC. 2004, 2: 1.

[13] Ratinov L, Roth D. Design challenges and misconceptions in named entity recognition[C]//Proceedings of the Thirteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2009: 147-155.

[14] Lin D, Wu X. Phrase clustering for discriminative learning[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics, 2009: 1030-1038.

[15] Passos A, Kumar V, McCallum A. Lexicon infused phrase embeddings for named entity resolution[J]. arXiv preprint arXiv:1404.5367, 2014.

[16] Suzuki J, Isozaki H, Nagata M. Learning condensed feature representations from large unsupervised data sets for supervised learning[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. Association for Computational Linguistics, 2011: 636-641.

大连理工大学学术学位硕士研究生学位论文中期报告评审意见表

学　号		学生姓名		导　师	
第一次中期检查 □			第二次中期检查 □		

导师考核意见（对学位论文内容完成情况、阶段性成果、论文进度、后续工作思路、预期目标、及论文工作存在的问题等进行考查，给出考核成绩和具体改进意见和建议）：

1） 考核成绩：□ 优秀，□ 良好，□ 中等，□ 及格，□ 不及格

2） 是否通过：□ 通过，□ 不通过

3） 关于开题报告撰写质量及学位论文工作的具体意见（可加页）：

导师签字：

年　　月　　日

<table>
<tr><td rowspan="8">评审专家组</td><td></td><td>姓名</td><td>职称</td><td>学科专业</td><td>是否博导</td><td>签字</td></tr>
<tr><td>组长</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="6">成员</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="7">专家组评审意见（对学位论文内容完成情况、阶段性成果是否正确、研究方案和进度是否合理、开题时方案是否需调整或已做了哪些调整，后续工作思路是否正确、工作进度是否有保障、预期目标能否实现、论文质量是否能够保证以及论文工作存在的问题等进行考查，给出考核成绩，投票表决是否通过，并给出具体改进意见和建议）：
1) 考核成绩：☐ 优秀，☐ 良好，☐ 中等，☐ 及格，☐ 不及格
2) 是否通过：☐ 通过，☐ 不通过
3) 具体意见（可以加页）：

组长签字：
年 月 日</td></tr>
<tr><td colspan="7">点长意见：

点长签字：
年 月 日</td></tr>
</table>