大连理工大学

学术学位硕士研究生学位论文开题报告

论文题目：面向新疆暴恐事件的命名实体识别和事件抽取研究

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 6 日

报告日期：2016 年 10 月 12 日

报告地点：大连理工大学创新园大厦报告厅

研究生院制表

说　明

学位论文开题考核是硕士研究生课程学习结束后开展学位论文工作的基本要求，是保证学位论文质量、工作进度和研究生培养质量的首要环节。

一、考核内容：首先，考查硕士生对本学科专业的基础理论与专业知识的掌握程度、课程学习情况等；其次，考查学位论文工作准备情况，包括论文选题、文献阅读、工作难度、研究思路、研究基础、写作能力和答辩表达能力等；此外，还要考查学术参与学术活动情况、及学习和工作态度等。

二、考核时间：原则上，硕士生的开题报告应在第 2 学期末进行。

三、报告撰写：开题报告正文字数不少于 6000 字；参考文献数量不少于 20 篇，其中，外文资料不少于二分之一，近 5 年文献不少于三分之一；正文及参考文献等撰写要求参见《大连理工大学硕士学位论文格式规范》。

四、考核办法：开题考核由学部（学院）按学科专业集中组织 3-5 名本学科领域专家以答辩的方式进行。硕士生进行口头陈述、答辩，研究生口头陈述时间不少于 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：开题报告一式两份，签字后分别由学部（学院）和学生保存。

六、信息登录：研究生开题后登录研究生信息管理系统上传开题报告（PDF 文档）及考核结果。

开题报告正文

1 课程学习情况（附成绩单）、参加科研和学术活动等情况

研究室第一学年基本修满对学术硕士研究生所要求的学分，目前已修学分29.0分，其中必修学分18.0分，选修学分11.0分，基本满足课业要求（剩余3分必修课在研究生第二学年内完成）。

课程名称	课程学分	选修学期	成绩
算法设计与分析	3	1	72
人工智能	2	2	94
分布式数据库	2	1	92
论文写作与学术规范	1	1	95
中国特色社会主义理论与实践研究	2	1	82
口语交流 I （基础口语表达）	1	1	84.5
阅读与写作 I （基础读写技能）	2	2	89
矩阵与数值分析	3	1	87
数理统计	2	1	86
数据仓库技术	2	2	P
搜索引擎与文本挖掘	2	1	86
分布式对象技术	2	2	P
中间件技术	2	2	P
自然辩证法概论	1	1	P
数据挖掘与知识管理	2	1	P

表 1.1 课程成绩

积极参加各类科研和学术活动，对和自己专业相关的专家讲座等活动积极参与，了解专业前沿知识，开拓自己的视野。同时也能积极参与实验室的学术学习和科研工作，实验室的一些项目，自己也都在能力范围内完成了部分工作。阅读了很多研究方向相关的论文，了解了研究方向的热点问题、难点问题，为自己下一步的学术研究找到了明确的方向。

2 学位论文研究背景、目的和意义

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，给两国社会和周边国家造成极大的冲击。

而我国境内新疆各地的稳定形势可以概括为：一是新疆社会大局稳定可控，二是稳定的基础依然脆弱，三是形势依然极其严峻、复杂；强调新疆反恐怖斗争已经进入比以往更加严峻复杂、更加尖锐激烈的新阶段，我们与“三股势力”的斗争是一场历史的较量。

但从国际形势看，很多风浪都与美国等西方国家策动的“颜色革命”不无干系。近年来，全球恐怖活动明显升温，尤其是“伊斯兰国”恐怖组织，建立政教合一的伊斯兰政权，煽动全世界包括新疆的穆斯林进行“圣战”。受特殊地缘、宗教、民族等因素的作用，新疆历来易受外来的影响，疆内出境参加暴恐组织的人员与境内勾连频发、拉拢煽动，刺激效应不断加剧。新疆发生的若干起恐怖活动，都是由境外恐怖组织直接指挥的。

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进行攻击。

在以上背景下，快速、实时地掌握国际媒体对新疆暴恐事件的情感倾向性，有助于开展积极的舆论斗争，及时分析研判国际上各种政治言论传播特点及渠道，掌握反恐斗争的主动权，打好主动仗。

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、事件抽取等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。因此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。

3 国内外研究现状及发展动态分析

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。

深度学习作为机器学习的一个分支，其自动学习特征表示得到了广大研究者的热烈追捧。由于传统的手工设计特征不仅花费大量的时间，而且设计的特征有时是只能应用于特定领域，加上有时可能设计者本身考虑不周全，设计出来的特征也不完整，更何况网络的发展，新词的不断出现，语言形式也在发展，每次更新都重新设计特征，这种循环设计是一种灾难。如果机器通过机器学习算法自动学习到特征表示，整个学习过程自动执行，那么很多任务就能被解决，至少节省大量的时间。深度学习作为特征学习方法提供了这样一种解决方案。

Bengio[1]的词向量为自然语言处理融合深度学习方法打开了一扇大门。在自然语言处理当中有一个经典的问题称为“维数灾难”，传统表示一个词的方式称为 one-hot 即词库多少个字，就有多少的维度，表示这个字的那一维为 1，其他为 0，也称为索引向量。据统计中国汉字有 8 万，常用的有 3500 字，牛津英语辞典单词量达 60 多万，英语六级也要会 5500 个单词，这么大的词汇使用索引向量表示相当的稀疏，训练模型时比较容易过拟合。在过去都采用手工设计特征和使用简单的线性模型作为目标函数共同解决这个问题。深度学习采用词向量来代替，它不仅低维度表示一个词，而且相似的词之间，词向量也能表示出它们的关联性，即词向量的这种关联性体现出词向量能在某种程度上表示语义，这是索引向量无法做到的。

命名实体识别(Named Entities Recognition, NER)是自然语言处理(Natural Language Processing, NLP)的一个基础任务。其目的是识别语料中人名、地名、组织机构名等命名实体，主要应用于信息抽取、信息检索、机器翻译、问答系统等领域。

NER 任务其本质是序列标注问题。其方法主要分为两类:基于规则的命名实体识别方法;基于统计机器学习的命名实体识别方法。

而在机器学习领域内，基于循环神经网络（Recurrent Neural Network,简称 RNN）的深度学习模型，在理论上克服了传统浅层模型（如条件随机场，CRF[2]）无法长远考虑上下文的问题，同时取消了传统神经网络输入相互独立的假设。其变种基于 LSTM[3]单元和 GRU[4]单元的 RNN 模型，更是解决了 RNN 训练时带来的梯度消失[5]的问题。

下面简单介绍近年来的相关研究进展。

Tjong Kim Sang[6]、Tjong Kim Sang 和 De Meulder[7]、Doddington[8]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[9]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[10]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[11]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[12]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[13]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，语义角色标注（Semantic Role Labeling）等 4 个经典问题。[14] 文献[13]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[15]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[16]采用了带有字符级 CNN 的 BRNN 进行关于德语的序列标注任务。

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。

事件抽取在研究中通常采用基于事件框架或基于本体的方法。Jiang B[22]等针对网络上的事件利用领域本体词、概念、关系等信息抽取事件的各种信息，取得不错的效果。研究人员针对新闻事件抽取出新闻中的 5W1H，通过这些信息构成一个新闻事件的框架，简单清楚、一目了然[23-24]。 Jakub Piskorski 等[25]利用聚类、语义分析等方法对互联网上的信息进行抽取，能够实时抽取到网络上的突发危机、自然灾害等事件。Atkinson M 等[26]对比几种事件抽取方法并提出了在网络中抽取事件的一些技巧，对研究人员有很大的帮助。

4 主要研究内容、研究目标、拟解决的关键问题

4.1 主要研究内容

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[18]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[17]进行词向量的训练。三、采用深度学习框架 Keras[21]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。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。

4.2 研究目标

本文研究目的是构建基于暴恐事件的命名实体识别模型。由于缺少关于新疆维文实体名在英文下的相关语法知识，在文献[18]的启发下，决定在词向量作为输入的基础上，结合字符向量来学习相关实体名的形态学信息，同时参考文献[13]考虑到标记之间存在相关一定的依赖关系，因而引入句级对数似然函数，并采用维特比算法进行预测，希望提高 NER 的 F1 值，尤其是和新疆维文相关的 NE 的 F1 值。采集国际媒体关于新疆暴恐事件的报道，抽取新闻事件 5W1H。

4.3 关键问题

本文拟解决的关键问题：一、相关暴恐事件的数据采集和预处理，由于 GDELT[19]汇集了全球各地国际新闻的来源，定时更新新闻，这对我们查找暴恐事件的报道提供了一条便捷的途径，通过 Google BigQuery[20]进行查询，获取链接后，使用爬虫技术进行新闻爬取工作，最后进行一定的处理，成为我们所需要的语料，这里的爬取和处理工作很重要，是之后工作的基础；二、利用英文维基百科进行词向量的训练，将下载英文维基百科上的语料，将 xml 格式转换为 text 后，使用 gensim 的 python 库进行词向量训练，训练的关键在于词向量的维数、窗口大小、最小出现次数的设定，一个好的词向量能对于结果有很大的影响；三、实现文献中提到的字符向量模型，这是本文中在缺乏维文实体名在英文中表达的语言知识情况下提出的解决方案，即利用形态学的特征，解决维文实体名在英文

下的识别问题；四、实现文献[13]中提出的优化句级对数似然函数，通过维特比算法，对序列进行预测，从而从全局的角度进行序列预测，提高 NER 的识别率。

5 学位论文的研究方法、技术路线、试验手段、关键技术等论述

5.1 研究方法

利用学校图书馆、网上相关学术数据库等资源，通过阅读命名实体识别、深度学习模型以及事件抽取的相关文献，提高自己的理论基础，并把握该方向最新动态，为下一步研究工作打下扎实的理论基础。根据相关文献，实现文献中的模型，在熟悉深度学习框架的同时，提高自己的编码调试能力。处理英文维基百科语料、调整词向量模型的参数以达到优化词向量训练效果。调整字符向量的模型参数，提升 NER 的准确率、召回率、F1 值。通过使用句级对数似然函数，提升 NER 系统的准确率、召回率、F1 值，并逐一进行对比，找出不足，进行微调。

5.2 技术路线

利用深度学习进行命名实体识别主要方法是把命名实体识别抽象成为给一个句子中所有的单词赋予一个标签的任务。具体步骤如下:

(1) 对语料进行标注

在新疆暴恐事件的命名实体识别过程中，我们采用 B (Begin，命名实体的开始)、I (Internal，命名实体的中间部分)、S (Single，代表该单词本身就是一个实体)、O(Other，其他)四个标注符号对每个语料进行标注。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, PERSON,B_ LOCATION, I_ LOCATION, E_ LOCATION, LOCATION, B_ORG, I_ ORG, E_ORG,ORG, O }，分别表示人名的开始、人名的中间、人名的结束、单独的人名，地名的开始、地名的中间、地名的结束、单独的地名，机构名的开始、机构名的中间、机构名的结束、单独的机构名，其他。

(2) 利用深度学习方法训练词向量

在深度学习的研究中，词向量一般都是用 Distributed Representation 表示的一种低维实数向量，通过这种表示方式，可以让相似的词的向量距离更加相近，因此我们使用经典的 Word2Vec 进行词向量的训练。

(3) 加入字符词向量

在 NER 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：

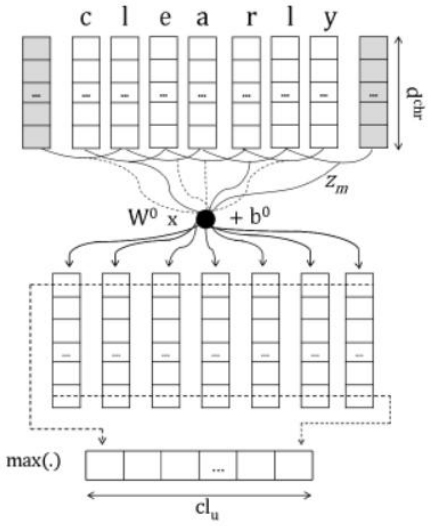

图 1.卷积方式获得字符向量

给定一个词 w，有 M 个字符，即 $w=\{c_1,c_2,...,c_M\}$,我们将每个字符 c_m 转换为对应字符向量 $r_m^{chr}\in W^{chr}$，其中 $W^{chr}\in\mathbb{R}^{d^{chr}\times|v^{chr}|}$，则有 $r^{chr}=W^{chr}v^c$，其中 v^c 采用的是 one-hot 编码。

卷积层的输入是字符向量，由字符序列构成，形如 $\{r_1^{chr},r_2^{chr},...,r_M^{chr}\}$，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 $z_m=(r^{chr}_{m-(k^{chr}-1)/2},...,r^{chr}_{m+(k^{chr}-1)/2})^T$

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（Max pooling）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了。

(4) 利用深度学习方法训练命名实体识别模型

命名实体识别中一个单词的标签主要受上下文前后几个词义的影响，通过滑动窗口方法获得一个词的上下文作为输入，利用模型进行训练得到该单词的输出标签。以滑动窗口大小是 n=5 为例，以单词 A 为中心取了 5 个单词作为输入，然后把这 5 个词在第二步得到的实数向量链接成一个向量，这个向量上面接若干隐含层，最后再接一个输出层。因为这个核心单词可能得到的标签有 13 个，即 13 分类问题，所以输出层为含 13 个节点的 softmax 层。该模型的参数为。$\theta=\{W^1,...,W^L\}$，对应的损失函数为交叉熵损失函数，再利用随机梯度下降法参数进行更新。

我们知道，一句话中，词与词之间是有语义联系的，给每个词贴上标签后，

它们之间也应该具有相关性，比如 B 后边不紧跟 0。为此，我们引入 CRF 层来表示这种相关性。它首先声明了一个转移矩阵 $A \in R^{|L|\times|L|}$，其中 $|L|$ 为标签数量。A_{ij} 表示从第 i 个标签转移到第 j 个标签的分数，而一句话中每个词生成一个的标签 $f_\theta(t_i \mid i)$，整个句子的得分 $s(w_{[1:n]}, t_{[1:n]}, \theta) = \sum_{i=1}^{n}(A_{t_{i-1}t_i} + f_\theta(t_i \mid i))$，目标函数 $\log p(t_{[1:n]} \mid w_{[1:n]}, \theta) = s(w_{[1:n]}, t_{[1:n]}, \theta) - \log(\sum_{\forall u_{[1:n]} \in T^n} e^{s(w_{[1:n]}, u_{[1:n]}, \theta)})$，用维特比算法找到 s 最大的一条路径。

新疆暴恐事件的抽取任务主要分三步：

(1) 新疆暴恐事件框架的抽取

下面的事件框架是基于对部分新疆暴恐事件新闻分析的基础上抽取出来的初步结果，框架中冒号前面的内容表示事件的侧面，冒号后面的内容是表示该侧面的侧面词。下面是初步抽取结果，侧面信息收集的并不全面，还有很多地方需要改进。

The Frame of Kunming Railway Station Violent Terrorist Case {
Perpetrators of Violent Terrorist Incident: Abdiryim Kurban
The Event Occurrence Time: The night of March 1，2014
Location of the Accident: Kunming Railway Station
Casualties: 31 died, 141 injured}

(2) 侧面信息抽取

对侧面信息的抽取主要使用基于规则的抽取方法，事件抽取规则是通过对语言规律(语法、语义等)的分析来获得一系列的句型规则来描述事件关键信息，从而将用户感兴趣的信息抽取出来。抽取规则的制订是一个重要的工作，因为事件的各个侧面的表示方法多种多样，而规则集的要求是尽可能覆盖到大部分的侧面信息。所以规则的制定要达到覆盖面广、精度高的要求。因此制定规则需要归纳总结训练语料中大量描述事件的句子。

例如下面描述暴恐事件中伤亡结果的规则: 1 {time} {in}2{place} {terrorist attacks} { about| at least}3{num}{ children| people| resident} { died| injured| missing } 规则中每一对大括号表示一个节点，括号前面的数字表示待抽取信息的编号，括号前没有编号的节点是关键词节点。从这条规则中可以看出编号 1 表示事件的发生时间，编号为 2 表示事件发生的地点，编号 3 为受伤、死亡等造成结果的数量词或短语。待抽取节点 time, place, num 分别表示时间短语、空间短语、数量词或短语。另外关键词节点中的“|”表示或的关系，即节点中的词有一个匹配成功则此节点匹配成功。

(3) 事件信息合并

事件侧面信息合并的实质是对抽取到的事件框架的不同侧面合并成一个统一侧面。众所周知，对于事件发生的地点应该判断地点间的是否有包含关系，将

粒度大的地点合并到粒度小的地点。对于其他侧面信息若信息之间发生冲突，则应该根据报道时间判断，信任报道时间靠后的新闻。因为通常报道时间越晚的新闻的数据可靠性越高，同一事件报道的结果也越详细。下面是侧面信息合并的具体算法流程。

```
输入：从不同报道中抽取出事件发生的时间、地点、结果
      等侧面信息后构成的事件侧面对象集合E
——initialize E'= null, i=0;
——for(event i in E){
Initialize j=i+1;
If (event i and event j represents the same info){
1: update event i and event j time area;
2: update event i and event j location area;
3: sort event i and event j by event time;
4: update E;
}else{
    Add event j to E'
}
    输出：合并同一事件后的事件对象集合E'。
```

图 2.事件抽取合并算法

5. 3 试验手段及关键技术

编写网络爬虫，采集一定数量的新闻数据，作为后续研究的语料。对英文维基百科的数据进行格式规整和数据清洗，满足 word2vec 的要求。预训练完的词向量作为模型的参数，将输入文本转化为词向量。采用带有字符向量的 RNN 模型进行训练，最后对序列采用维特比算法进行预测。计算本 NER 系统的准确率、召回率和 F1 值，根据实际实验结果进行相应调整，以期达到最佳。

6 年度研究计划

时间	任务
2016 年 09 月-2016 年 10 月	阅读文献和查阅资料
2016 年 11 月	方案设计、语料采集和预处理
2016 年 12 月	配置词向量训练工作环境，语料标注
2017 年 01 月	实现字符向量模块，实现句级对数似然函数
2017 年 03 月	进行实验，训练和测试，优化 NER 模型
2017 年 04 月--2017 年 10 月	完成事件抽取
2017 年 11 月--2018 年 03 月	进行系统测试
2018 年 04 月	完成论文初稿
2018 年 05 月	完成论文终稿

7 现有的研究基础

(1) 已阅读一定的文献，具备了本研究方向的相关理论基础

(2) 已经掌握文本处理技术，熟练使用 Python 等工具进行数据爬取与清洗。

(3) 已基本掌握 Keras[21]深度学习架构，具备一定的编程能力。

(4) 掌握基本的深度学习算法尤其是 RNN 模型的相关知识，并提前完成了 Word2Vec 训练词向量部分的工作。

8 现有研究条件及可能遇到的困难和问题分析

8.1 现有研究条件

一是已完成词向量训练环境的搭建以及训练。

二是已经从 GDELT 上爬取 6000 篇左右的新闻报道。

三是已完成基于英文维基百科语料的词向量训练。

四是团队支持：整个实验室中一共有 2 位同学进行相关的研究，不定期的思想和研究成果交流，能促进思维的扩散，为研究提供各种帮助。同时，定期的组会交流，有助于加速科研的进程。

五是学校提供的学术资源数据库，方便文献的查阅。

8.2 可能遇到的困难

对于 CRF 层的理解不够深入，一时间可能无法实现，需要更深入的阅读文献。此外，深度学习框架 Keras 的使用还仅仅停留在模块调用，还无法独立设计，可能会影响任务进度，需要深入阅读框架代码，为下面的实验做准备。

8.3 问题分析

通过查阅相关文献、广泛参考相关实验数据，并根据已有实验结果调整参数。

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] 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-ww289.

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

[4] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[5] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE transactions on neural networks, 1994, 5(2): 157-166.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.
[14]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.
[15]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.
[16]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.
[17]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.
[18]dos Santos C N, Zadrozny B. Learning Character-level Representations for Part-of-Speech Tagging[C]//ICML. 2014: 1818-1826.
[19]Leetaru K, Schrodt P A. Gdelt: Global data on events, location, and tone, 1979–2012[C]//ISA Annual Convention. 2013, 2(4).
[20]Google Corp. Google BigQuery.[EB/OL]. (2011-11-14) [2016-10-07] https://bigquery.cloud.google.com/
[21]Chollet F. Keras[J]. GitHub repository: https://github.com/fchollet/keras, 2015
[22]Jiang B, Zhu M, Wang J. Ontology-Based Information Extraction of Crop Diseases on Chinese Web Pages[J]. Journal of computers, 2013, 8(1): 85-90.
[23]Wang W. Chinese news event SW1H semantic elements extraction for event ontology population[C]//Proceedings of the 21st international conference companion on World Wide Web. ACM, 2012: 197-202.
[24]Tung C M, Lu W H. Detect Negative Event for Depression Tendency from Web Blogs[C]//The 15th International Conference on Biomedical Engineering. Springer International Publishing, 2014: 801-804.
[25]Piskorski J, Tanev H, Atkinson M, et al. Online news event extraction for global Springer crisis surveillance[M]//Transactions Heidelberg,2011: 182-212. on computational collective intelligence V. Berlin
[26]Atkinson M, Du M, Piskorski J, et al. Techniques for Multilingual Security-Related Event Extraction from Online News[M]//Computational Linguistics. Springer Berlin Heidelberg, 2013: 163-186.