

**学术学位硕士研究生学位论文开题报告**

论文题目： 基于新疆暴恐事件的命名实体识别研究

|  |  |
| --- | --- |
| 姓 名： | 林广和 |
| 学 号： | 21509174 |
| 学科专业： | 计算机科学与技术 |
| 指导教师： | 张邵武副教授 |
| 入学日期： | 2015年9月6日 |
| 报告日期： | 2016年10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

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（事件抽取、命名实体识别、信息时代、）

随着信息的爆炸式增长，一个大规模生产、分享和应用数据的时代正在开启。 《纽约时报》在2012年的一篇专栏中也指出:“大数据时代己经来临，在商业、经济、传媒及其他领域中，决策将日益基于数据和分析而做出，而并非基于经验和直觉”。[[1]](#endnote-1)

这表明，数字洪流正以前所未有的铺天盖地之势向我们涌来，无处不在地渗入到人们生活的各个层面。无论你是哼着小曲，迈着愉快的步子走过公司的摄像头，进入办公室，还是在餐馆用IPAD点喜欢吃的菜肴，亦或是下班了跟朋友在微信上分享自己的心情，不管采取任何一种行为，人们的生活都在无形中被数字影响着。在大数据时代，这一切都将作为数据资料被记录下来。

大数据时代下的数据爆炸一方面加大了人们选择信息的空间，另一方也对传统媒体和读者提出了更大的挑战。网络新闻的海量性完全冲破了传统媒体新闻在数量上的控制和局限，它几近把一个完整的大千世界活灵活现地呈现在人们的面前。读者面对的再也不是单调、简单的几家几条新闻，而是丰富的、多元的、开放的和全面生动的新闻世界。网络新闻的海量性加大、加宽了人们接受新闻的选择余地，提高了其能动性。

正因人们被包围在信息的汪洋大海之中 ,如何准确有效的从大量 无序、杂乱、无结构的信息中提取感兴趣的事件信息已成为亟待解决的问题, 事件抽取技术正是解决上述问题的有力工具之一 。事件抽取(Event Extraction)隶属于信息抽取领域，主要研究如何从含有事件信息的非结构化文本中抽取出用户感兴趣的事件信息, 用自然语言表达的事件以结构化的形式呈现出来，如什么人, 在什么地方 ,什么时间,做了什么事等。[[2]](#endnote-2)

而要想正确的获取人名、地名、时间等实体信息，命名实体识别是事件抽取过程中一项不可或缺的技术。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。

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[[3]](#endnote-3) 的词向量（检查一下这句话，是否符合实际情况）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。

RNN->NER

新疆暴恐事件引入

~~在自然语言处理领域具有代表性的深度学习模型就是Google公司提出的word2vec模型。Word2vec通过训练，可以把对文本内容的处理简化为K维向量空间中的向量运算，而向量空间上的相似度可以用来表示文本语义上的相似度。因此，word2vec输出的词向量是很多NLP相关的基础工作，比如词性标注（POS）、NER等。~~

3国内外研究现状及发展动态分析（命名实体识别：传统的统计学习方法、深度学习；事件抽取）

Ratinov and Roth (2009) [[4]](#endnote-4) 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。

（Ratinov and Roth (2009) used non-local features, a gazetteer extracted from Wikipedia, and Brown-cluster-like word representations, and achieved an F1 score of 90.80 on CoNLL-2003.）

Lin and Wu (2009)[[5]](#endnote-5) 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 等人(2014)[[6]](#endnote-6) 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等人 (2011)[[7]](#endnote-7)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 等人[[8]](#endnote-8) 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。[[9]](#endnote-9) 文献8 在NER训练时采用了句级似然函数，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（2015）[[10]](#endnote-10)提出了CharWNN的网络，该网络是对Collobert 等人8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（2015）[[11]](#endnote-11)采用了带有字符级CNN的BRNN进行关于德语的序列标注任务。

~~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等~~ ~~[[12]](#endnote-12)~~~~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~~~~[[13]](#endnote-13)[[14]](#endnote-14)~~~~。Jakub Piskorski 等~~~~[[15]](#endnote-15)~~~~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~~~~[[16]](#endnote-16)~~~~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（具体解决问题的几个步骤阐述）

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。然而，到目前为止，很少有应用深度学习的NER系统，采用字符级向量。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gensim的python库进行词向量的训练。三、采用深度学习框架Keras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。

（缺少 事件抽取）

4.2 研究目标(NER和EE的实验效果)

本文研究目的是构建基于暴恐事件的命名实体识别模型~~事件抽取研究的模型~~。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[[17]](#endnote-17)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8 考虑到标记之间存在相关一定的依赖关系，因而引入了句级似然函数，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。

（缺少 事件抽取）

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[[18]](#endnote-18) 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[[19]](#endnote-19)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17 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，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8 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。

缺少 事件抽取

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，改进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:

第一步、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。

第二步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。

第三步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。，对应的损失函数如公式1，其中NLL为negative log likelihood的缩写，N为所有数据集中单词的个数。再利用随机梯度下降法对进行更新。

缺少 事件抽取

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。预训练完的词向量做为模型的参数，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 |
| 2017年01月 | 实现字符向量模块 |
| 2017年03月 | 将RNN模型融合字符向量进行实验 |
| 2017年4月—2016年06月 | 实现句级对数似然函数 |
| 2017年07月—2017年10月 | 训练和测试，优化模型结构 |
| 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[[20]](#endnote-20)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

使用词向量表示短文本的模型不能完全表达句子的语义。其原因可能有几个方面：一是采集的短文本数量和质量没有达到Word2vec的要求；二是文本中词语的上下文、词性和词序的影响；三是一词多义的影响。要不断加强深度学习的理论学习，研究国内外最新的研究成果，能够快速发现问题原因并解决。

对于字符向量的理解不够深入，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。

9参考文献

[1] Hinton G E. Learning distributed representations of concepts[C]//Proceedings of the eighth annual conference of the cognitive science society. 1986, 1: 12.

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

[3] 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.

[4] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.

[5] Zhang M, Zhang Y, Che W, et al. Chinese Parsing Exploiting Characters[C]//ACL (1). 2013: 125-134.

[6] Sun Y, Lin L, Yang N, et al. Radical-Enhanced Chinese Character Embedding[C]//Neural Information Processing. Springer International Publishing, 2014: 279-286.

[7] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. 2001.

[8] Socher R, Lin C C, Manning C, et al. Parsing natural scenes and natural language with recursive neural networks[C]//Proceedings of the 28th international conference on machine learning (ICML-11). 2011: 129-136.

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

[10] Kusner M J, Sun E D U Y, Kolkin E D U N I, et al. From Word Embeddings To Document Distances[J]，2015.

[11] Lai S, Liu K, Xu L, et al. How to Generate a Good Word Embedding?[J]. arXiv preprint arXiv:1507.05523, 2015.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.

[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] 孙志军, 薛磊, 许阳明, 等. 深度学习研究综述[J]. 计算机应用研究, 2012, 29(8): 2806-2810.

[16] Rubner Y, Tomasi C, Guibas L J. The earth mover's distance as a metric for image retrieval[J]. International journal of computer vision, 2000, 40(2): 99-121.

[17] Salton G, McGill M J. Introduction to modern information retrieval[J]. 1986.

[18]龚静, 周经野. 一种基于多重因子加权的文本特征项权值计算方法[J]. 计算机技术与自动化, 2007, 26(1): 81-83.

[19]杨阳, 刘龙飞, 魏现辉,等. 基于词向量的情感新词发现方法[J]. 山东大学学报：理学版, 2014(11):51-58.

[20]Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1988, 24(5): 513-523.

[21]Ronan Collobert, Jason Weston. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. ICML(2008).

**大连理工大学学术学位硕士研究生学位论文开题报告评审意见表**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 学 号 | |  | | 学生姓名 | |  | 导 师 | |  | |
| 第一次开题 □ | | | | | | 第二次开题 □ | | | | |
| 导师考核意见（对课程学习情况、参加学术活动情况、学位论文工作及开题报告撰写情况、学习和工作态度等进行考查，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 | | | | | | | | | | |
| 评  审  专  家  组 |  | 姓名 | 职称 | | 学科专业 | | | 是否博导 | | 签字 |
| 组长 |  |  | |  | | |  | |  |
| 成员 |  |  | |  | | |  | |  |
|  |  | |  | | |  | |  |
|  |  | |  | | |  | |  |
|  |  | |  | | |  | |  |
|  |  | |  | | |  | |  |
|  |  | |  | | |  | |  |
| 专家组评审意见（对课程学习情况、参加学术活动情况、学位论文工作及开题报告撰写情况、学习和工作态度等进行考查，给出考核成绩和学位论文工作具体改进意见和建议）：   1. 选题是否属于本学科领域（含交叉学科）：□ 是，□ 不是（须重新开题） 2. 考核成绩：□ 优秀，□ 良好，□ 中等，□ 及格，□ 不及格 3. 是否通过：□ 通过，□ 不通过 4. 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）：     组长签字：  年 月 日 | | | | | | | | | | |
| 点长意见：  点长签字：  年 月 日 | | | | | | | | | | |

1. 安邦.把握住大数据时代的企业才能引领未来.[EB/OL]. (2012-11-15). [2016-10 06].httn://bloe.sina.com.cn/s/bloea6d8ee200101bonb.html. [↑](#endnote-ref-1)
2. 赵妍妍.中文事件抽取的相关技术研究[ D] .哈尔滨:哈尔滨工业大学, 2007. [↑](#endnote-ref-2)
3. Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research, 2003, 3:1137-1155. [↑](#endnote-ref-3)
4. Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning, pages 147–155. Association for Computational Linguistics. [↑](#endnote-ref-4)
5. Dekang Lin and Xiaoyun Wu. 2009. Phrase clustering for discriminative learning. In 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, pages 1030–1038. Association for Computational Linguistics. [↑](#endnote-ref-5)
6. Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning, pages 78–86. Association for Computational Linguistics. [↑](#endnote-ref-6)
7. Jun Suzuki, Hideki Isozaki, and Masaaki Nagata. 2011. Learning condensed feature representations from large unsupervised data sets for supervised learning. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers, pages 636–641. Association for Computational Linguistics. [↑](#endnote-ref-7)
8. Ronan Collobert, Jason Weston, L´eon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011b. Natural language processing (almost) from scratch. The Journal of Machine Learning Research, 12:2493–2537. [↑](#endnote-ref-8)
9. 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. [↑](#endnote-ref-9)
10. Cıcero Santos, Victor Guimaraes, RJ Niter´oi, and Rio de Janeiro. 2015. Boosting named entity recognition with neural character embeddings. In Proceedings of the Fifth Named Entities Workshop, pages 25–33. [↑](#endnote-ref-10)
11. Matthieu Labeau, Kevin L¨oser, and Alexandre Allauzen. 2015. Non-lexical neural architecture for fine-grained POS tagging. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 232–237. Association for Computational Linguistics. [↑](#endnote-ref-11)
12. 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. [↑](#endnote-ref-12)
13. Wang W. Chinese news event 5W1H semantic elements extraction for event ontology population[C]//Proceedings of the 21st international conference companion on World Wide Web. ACM, 2012: 197-202. [↑](#endnote-ref-13)
14. 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. [↑](#endnote-ref-14)
15. Piskorski J, Tanev H, Atkinson M, et al. Online news event extraction for global crisis surveillance[M]//Transactions on computational collective intelligence V. Springer Berlin Heidelberg, 2011: 182-212. [↑](#endnote-ref-15)
16. 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. [↑](#endnote-ref-16)
17. 关于字符向量的文献待插入 [↑](#endnote-ref-17)
18. Leetaru K, Schrodt P A. Gdelt: Global data on events, location, and tone, 1979–2012[C]//ISA Annual Convention. 2013, 2(4). [↑](#endnote-ref-18)
19. Google Corp. Google BigQuery.[EB/OL]. (2011-11-14) [2016-10-07] <https://bigquery.cloud.google.com/> [↑](#endnote-ref-19)
20. Chollet F. Keras[J]. GitHub repository: https://github.com/fchollet/keras, 2015. [↑](#endnote-ref-20)