

**学术学位硕士研究生学位论文开题报告**

论文题目： 基于暴恐事件的事件抽取研究

|  |  |
| --- | --- |
| 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，Bengio[[3]](#endnote-3) 的词向量为深度学习进入自然语言处理打开了一扇大门。它在训练语言模型的过程中将词映射到一个更抽象的低维向量空间中，解决了传统的稀疏表示法带来的维度灾难问题。

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）

在自然语言处理当中，词的表示方法是基础技术之一。很多学者采用独热编码（One-Hot Encoding）将词表示成一个one-hot的向量，其长度与词汇的数量一样，只有该词的维度为1，其他维度都为0。one-hot表示方法没有考虑词与词之间的语义相关性，词与词之间相互独立。由于短文本长度很短，该问题就更加凸显，如短文本“联想笔记本”和“惠普台式机”，如果二者利用one-hot的表示方法其相似度近似为0。但是二者文本意义却存在相关性，同样都是与电脑相关。为了解决这个问题，很多学者不断研究，并取得了一定的成果。随着深度学习的发展，许多学者主要集中在如何利用一个连续的实数型向量来表示词，也就是本文所说的词向量模型。词向量模型的目标是将词映射到一个低维的实数型向量空间当中，使得词义越相近的词在空间上的距离越近。

1986，Hinton等人[1]首先提出了词向量的概念，利用它来解决文章中“维度灾难”的问题。2003年，Bengio等人[2]提出了基于前馈神经网络的语言模型（NNLM），该语言利用句子前面出现的词作为上下文信息来预测下一个单词。然而该模型的时间复杂度非常高，训练速度非常慢。Collobert、Weston等人[3]提出了C﹠W神经网络语言模型来学习词向量，该模型利用排序（Ranking-Type）的损失函数。2013年，谷歌Google公司的Mikolov等人[4]提出并开源了语言建模工具 word2vec，并在自然语言处理领域得到了广泛关注。word2vec模型实现了连续bag-of-word模型[12]，以及计算词向量的 skip-gram 结构[13]。它以文本集为输入，通过训练生成每个词对应的词向量。这些词向量可以作为词的特征应用到其他自然语言处理问题中去。比如可以根据词向量计算两个词的相似程度，使用词向量可以避免词语表示的“维灾难”现象。由于word2vec模型简单高效，受到了学术界和工业界的广泛关注。

基于词向量的自然语言处理方法已经取得了非常大的成功。Collbert和Weston等人[3]利用词向量特征进行词类标注、命名实体识别等。Zhang[5]和Sun[6]等人改进词向量学习方法，并将其成功应用到中文分词当中，其分类准确率比人工大量特征挖掘的条件随机场效果CRF(Conditional Random Field) [7]还要好。杨阳等人[19]将其应用到情感新词的发现。

这些研究均集中在词级别上，在文档级别的研究仍不是太多。ICML2011（ICML即国际机器学习大会），[R Socher](http://scholar.glgoo.org/citations?user=FaOcyfMAAAAJ&hl=zh-CN&oi=sra)、[CC Lin](http://scholar.glgoo.org/citations?user=u1BA6gkAAAAJ&hl=zh-CN&oi=sra)等人[8]在论文中提到利用根据句法树加权的方式来表示句子。ICML 2014，[QV Le](http://scholar.glgoo.org/citations?user=vfT6-XIAAAAJ&hl=zh-CN&oi=sra)、[T Mikolov](http://scholar.glgoo.org/citations?user=oBu8kMMAAAAJ&hl=zh-CN&oi=sra)等人[9]在论文《Distributed Representations of Sentences and Documents》中提出了doc2vec的概念，在word2vec模型上增加一个段落向量。该模型和Word2vec类似，也存在两种模型：DM 试图在给定上下文和段落向量的情况下预测单词的概率，DBOW 则在仅给定段落向量的情况下预测段落中一组随机单词的概率。ICML2015，MJ Kusner、EDUY Sun等人[10]在论文新提出一种计算文档相似度的方式，引入了WMD的概念，将词之间的余弦距离作为地表距离(ground distance)，求地球移动距离的线性规划最优解。

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 and Roth。

（Lin and Wu (2009) surpassed them without

using a gazetteer by instead using phrase features

obtained by performing k-means clustering over a

private database of search engine query logs.）

Passos et al. (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。

（Passos et al. (2014) obtained nearly the same performance

using only public data by training phrase vectors in

their lexicon-infused skip-gram model.）

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。

（ In order to

combat the problem of sparse features, Suzuki et al.

(2011) employed large-scale unlabelled data to perform

feature reduction and achieved an F1 score of

91.02 on CoNLL-2003, which is the current state of

the art for systems without external knowledge.）

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）

目前利用词向量表示短文本的方法，只要是将短文本里出现的词的词向量简单的叠加，并没有考虑词序带来的影响以及词权重不同的问题，在短文本聚类上也没有考虑在利用词向量表示句子的基础上进行，我的研究内容主要包括以下四个方面：一是短文本语料的准备，通过爬虫技术获取一定规模的短文本数据，并对爬取的短文本数据进行分词、去停用词等预处理；二是利用Word2vec训练预处理后的短文本语料，将单词表示成低维的词向量；三是利用Word2vec的训练结果，将短文本表示为词向量的集合，在前人的基础上增加词的权重和词序等因素，使其更能体系原句子的结构和主题；四是利用类词向量的短文本表示结果，选择合适的距离度量函数，计算不同短文本之间的距离对短文本进行聚类。

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)

本文的研究目标是以实验室高性能计算平台为基础，搭建Word2vec训练模型，并通过训练得到的词向量，在特征词权重的约束下使用词向量表示短文本，通过新的距离度量函数，计算两个短文本词向量集合的距离，在此基础上实现短文本的聚类，提高短文本聚类的准确率、召回率和F值。

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，一是短文本数据的采集和预处理。使用Word2vec进行词向量训练，训练选择的模型、语料、参数三个方面都会影响词向量的结果。[S Lai](http://scholar.glgoo.org/citations?user=HugytnQAAAAJ&hl=zh-CN&oi=sra), [K Liu](http://scholar.glgoo.org/citations?user=DtZCfl0AAAAJ&hl=zh-CN&oi=sra), L Xu等人[11]在论文中证明小规模的领域内语料比大规模的领域外语料训练的结果更好。选择合适的领域和合适规模的语料对于词向量的训练至关重要。二是Word2vec模型的参数选择，合理选择向量维度、上下文窗口大小、高频词亚采样的阀值、负例数目、被截断的低频词阀值的数量和是否采用层次softmax、是否使用CBOW算法[21]等选项的选择，选择合适的参数，以保证得到最佳的训练和测试效果；三是要构造好合适的短文本表示方法，即利用好Word2vec的训练结果，又综合考虑短文本的特点。四是要构造合适的集合距离度量函数，即能准确的反应不同短文本之间的相似度，又要保证该模型的计算量要在实验室计算平台的处理能力范围之内，以保证最终短文本聚类的实时性。

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

（1）分析目前常用短文本聚类技术的原理、具体应用和优缺点。

（2）分析目前利用词向量表示短文本的实现方法，准确度和实用性。

（3）重点分析词向量的原理和特点，以及在短文本聚类的可行性。

（4）利用词权重计算方法，加入词权重对句子语义的影响。同时考虑单词在句子中位置、长度以及词共现，提出利用词向量表示短文本的新方法，使其更贴近原句的。

（5）引入EMD距离[16]作为集合距离度量函数，计算不同短文本之间的相似度，并将这个距离度量应用于聚类算法中，从而实现文本聚类，即能准确的反应不同，又要保证该模型的计算量要在实验室计算平台的处理能力范围之内。

（6）将本文提出的模型与一些经典的文本聚类模型进行对比分析，提出本文模型的优缺点及其在实际应用中存在的问题和需要改进的地方。

（7）整理文献资料、代码和数据等，完成论文。

基于深度学习的新疆暴恐事件的命名实体识别

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

编写网站爬虫，采集一定数量的短文本数据，作为后续分析的语料。对爬取的文本数据进行格式规整和数据清洗，在此基础上进行中文分词、去停用词等预处理，使得符合Word2vec模型的输入要求。在高性能计算平台上对预处理后的语料进行训练，将输入的文本转化为词向量。利用词向量结果进行短文本的信息表示，构造距离度量函数，进行短文本聚类测试。计算短文本聚类结果的准确率、召回率、F值，根据测试结果进一步修改短文本表示方法和距离度量函数。

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。

5.4关键技术

5.4.1 词权重计算

将短文本中出现的词进行权重计算。因为不同的词具有的文本分类能力是不同的，对于含有较多文本分类信息，应该赋予较高的权重，相反为分类提供较少信息的词应该赋予较少权重。特征词权重计算对文本分类结果的准确性有很大影响，我们这里将词的TF-IDF值[17]、词序、词长以及词共现等因素综合考虑，计算不同词的权重，我们这里提出一个新的词权重的计算方法WWC系数（Word Weight Coefficient），公式为：



 TF-IDF（term frequency–inverse document frequency）是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。1973年，Salton[17]提出了TFIDF算法，并详细阐述了词权重在文献检索时的适用情况[20]。

TFIDF 权重计算公式如下： 。

其中词频（term frequency，TF）指的是某一个给定的词语在该文件中出现的频率。它的计算公式为：。

逆向文件频率（inverse document frequency，IDF）是一个词语普遍重要性的度量。它的计算公式为： 。

龚静[18]证明词语在文本的不同位置上所具有的意义是不同的，不同位置的特征词，代表文本内容的程度不同。加入位置信息后的特征词权重计算公式：



上式中 为位置权重因子，为特征词在文本中的权重，为特征词在文本相应位置出现的频率，令。

一般来说词语的长短与其出现的频率成反比，即词语越长则出现的次数越少。较短的词语出现的频率高，包含更多的信息，较长的词语出现频率小，但表达的内容更加具体。增加长词语的权重，可以更加准确反映特征词在文本中的重要性。加入词长权重的特征词权重计算公式为： ,上式中a表示特征词 的长度，令。

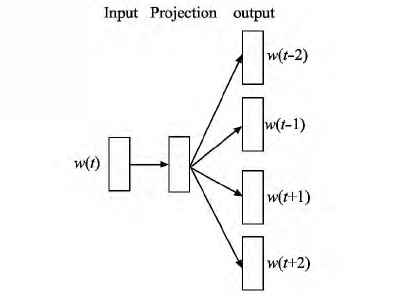
在中文文本分类中，词语的同现是词与词之间最直接的关系。在一个词中出现的所有字，两两之间的同现关系是词的意思所在[39]。词在文本中出现的频率为，词频为，词在文本中出现的频率为，词频为，词和同现频率为，则=。词和同出现的概率为，

公式为：

对于文本中所有词与词的同现概率可以组成一个空间概率矩阵，这个矩阵为一个n列的对称矩阵，n表示文本特征词的维数。用此矩阵对特征词在文档中的权重进行修正，公式如下：，令

5.4.2 Skip-gram模型

Skip-gram模型是Google公司在2013年发布的开源工具word2ve中的语言模型，用于从大规模非标注的文本数据中学习单词的分布式词向量。其本质特征在于利用滑动窗口捕捉词语的共现信息，为每个词生成高维的分布式词向量，使得所生成的词向量具有词语之间的语义和语法信息。

****

同时，Skip-gram中的每个词向量表征了上下文的分布。Skip-gram中的skip是指在一定窗口内的词两两都会计算概率，就算他们之间隔着一些词，这样的好处是“白色汽车”和“白色的汽车”很容易被识别为相同的短语。

5.4.2 EMD距离

EMD距离是在2000年由YOSSI RUBNER[16]在提出的并应用于计算视觉问题上，它可以被理解为从一种分布变换为列一种分布的最小代价。我们在这里引入EMD距离的概念，将词向量表示的短文本作为一种分布，计算它到列一个词向量表示的短文本的距离，判断不同短文本之间的文本相似度。设

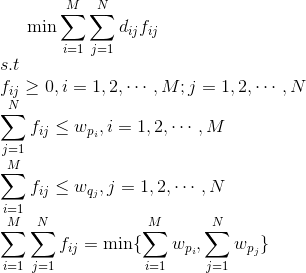
http://img.blog.csdn.net/20140509160644765

http://img.blog.csdn.net/20140509160740265

其中 http://img.blog.csdn.net/20140509152352421，

代表一个短文本的特征，代表另一个短文本的特征，代表特征的权重，代表特征的权重，代表和之间的距离，

通过以下最优化问题：



解出，EMD的定义为：

。

将该距离作为两个短文本之间的距离，并应用到聚类算法中。

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。

二是已完成CNKI论文信息爬取工具的研发。已经完成CNKI爬虫的开发，目前已实现自定义搜索关键字的论文信息提取。目前按照学校授予单位为关键词，已爬取“武书连2015中国大学排行榜”前50位+艺术类排名前5位的高校硕博士论文部分信息,数据总计52.8 MB，共包含259952条论文信息。

三是已完成现有CNKI论文信息集的词向量训练。

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。

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).

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). [2015-03 12].httn://bloe.sina.com.cn/s/bloe a6d8ee200101bonb.html.**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/bloe a6d8ee200101bonb.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]. (日期?) [2016-10-07] <https://bigquery.cloud.google.com/> [↑](#endnote-ref-19)