****



**研 究 生 毕 业 论 文**

**（申请硕士学位）**

**论文题目 基于卷积神经网络的实体关系抽取研究**

**作者姓名** 王强

**学科、专业名称**  **计算机科学与技术系**

**研 究 方 向 数据挖掘**

**指导教师** **李宁 副教授**

**2017 年 5 月 27 日**

**学号：MG1333075**

**论文答辩日期：2017 年 5 月 27日**

**指 导 教 师： （签字）**

|  |  |  |
| --- | --- | --- |
| **NJU2** | | |
| 南京大学申请硕士学位论文 | | |
| 基于卷积神经网络的实体关系抽取研究 | | |
|  | | |
| 作 者： | **王强** | |
| 专 业： | **计算机科学与技术** | |
| 研究方向： | **数据挖掘** | |
| 指导教师： | **李宁 副教授** | |
|  | | |
| **南京大学计算机科学与技术系**  **2017年5月** | | |
| **NJU2** | | |
|  | | |
| Research on Entity Relation Extraction Based on Convolutional Neural Network | | |
|  | | |
| Presented By  **Wang Qiang** | |
| Supervised by  **Prof. Li Ning** | | |
| A DISSERTATION  FOR THE APPLICATION OF MASTER DEGREE  SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE  AND TECHONOLOGY OF NANJING UNIVERSITY | | |
| **May 2017** | | |

声明

本人声明所呈交的论文是我个人在导师指导下、在南京大学及导师提供的研究环境（含标明的项目资助）下作为导师领导的项目组项目整体的组成部分而完成的研究工作及取得的研究成果。

除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

南京大学及导师所有权保留：送交论文的复印件，允许论文被查阅和借阅；公布论文的全部或部分内容；可以采用影印、缩印或其它复制手段保存该论文。

学生签名： 日期：

导师签名： 日期：

Declaration

I make a declaration here that the thesis submitted is composed of the researching work by myself and its corresponding researching results finished as a constituent part of the whole project in the project team lead by my advisor. The thesis is completed with the guidance of my advisor, and under the researching circumstances offered by Nanjing University and my advisor (including the project support indicated).

The thesis does not include other people’s researching results ever published or composed, except that are specially annotated and acknowledged somewhere in the article. Any contribution made to the research by my working partners is declared explicitly and acknowledged in the thesis.

Nanjing University and the advisor retain the copyright as follows: submitting the copies of the thesis, allowing the thesis to be consulted and borrowed; publicizing the whole or part of the thesis’ content; keeping the thesis by photocopy, microcopy or other copy methods.

Author Signature： Date：

Advisor Signature： Date：

摘要

随着信息技术的快速发展，互联网上产生了大量的非结构化的文本数据，其中包括：新闻，博客，政府文档，聊天日志等。如何帮助人们快速地从中获取有价值的信息成为计算机相关领域学者所关注的问题。实体关系抽取正是在这种背景下产生的，它的主要任务是识别文本中的实体并抽取实体之间的语义关系。现有的有监督学习关系抽取方法已经取得了较好的效果，但它们严重依赖词性标注、句法分析等自然语言处理工具提供分类特征。而这些工具往往存在大量错误，这些错误将会在关系抽取系统中不断传播放大，最终影响关系抽取的效果。为了避免过多依靠复杂的特征工程，本文提出利用卷积神经网络来解决关系抽取问题，其将会自动从句子中学习特征，以构建句子的分布式表示作为关系分类模型的输入，从而最小化对一些NLP处理工具和资源的依赖。并在此基础上设计并构建了面向互联网新闻文本的企业实体关系抽取系统。

本文的工作主要包括以下几个方面：

1)针对互联网新闻当中存在的企业实体关系，本文提出了基于Bootstrapping技术构建关系语料库的方法，该方法克服了纯人工标注过程中费时费力的缺点；

2)针对传统词袋模型在表征句子时缺乏语义信息以及未考虑词的位置信息的缺陷，本文提出了基于词向量加权和基于卷积神经网络的方法用于构建紧凑且具有语义的句子的分布式表示，作为关系分类模型的输入；

3)在前几部分的工作基础上，结合网页正文提取、命名实体识别等关键技术构建了企业关系抽取系统的原型。

**关键词**： 实体关系抽取；卷积神经网络；词向量；Bootstrapping

Abstract

With the rapid development of information technology, the Internet has produced a large number of unstructured text data, including news, blog, government documents, chat logs, etc. How to help people get valuable information from the web quickly becomes the problem concerned by the scholars of computer science areas. Entity Relation Extraction is generated in this context. Its main task is to identify the entities in the text and extract the semantic relations between the entities. The existing method based on supervised learning has achieved good results, but they rely heavily on POS, syntactic parsing, and other natural language processing tools to provide classification features. And these tools tend to have a lot of errors. These errors will continue to propagate in the relationship extraction system. In order to avoid relying on complex feature engineering, this paper proposes to use the convolution neural network to solve the problem. It will automatically learn features from the sentence to construct the sentence's Distributed representation as the input of the relational classification model. Thereby minimizing the reliance on some NLP processing tools and resources. And design and build a company relationship extraction system for Internet news text.

The main work of this paper is as follow：

1. In view of the existence of the company entity relation showed in the Internet News, this paper proposes a method of constructing relational corpus based on Bootstrapping technology. This method overcomes the shortcomings of time consuming and laborious in the process of manual labeling.

2. In the case of the shortcoming of the traditional word bag model lacks semantic information and the location information of the word when representing sentences, this paper propose two methods based on weighting of Word Embedding and Convolutional Neural Network to construct a compact and semantic Distributed representation of the sentences. As the input of the relational classification model.

3. On the basis of the first part of the work, combining the key technology of web page content extraction and Named Entity Recognition, this paper constructs the prototype of enterprise relationship extraction system.

**Keywords**: Entity Relation Extraction; Convolutional Neural Network; Word Embedding;

Bootstrapping

目 录

[第一章 绪论 4](#_Toc480974737)

[1.1 研究背景及意义 4](#_Toc480974738)

[1.2 关系抽取研究现状 5](#_Toc480974739)

[1.3 本文主要工作 7](#_Toc480974740)

[1.4 本文组织结构 8](#_Toc480974741)

[第二章 关系抽取的相关研究 9](#_Toc480974742)

[2.1 基于半监督学习的关系抽取 9](#_Toc480974743)

[2.1.1 DIPRE 10](#_Toc480974744)

[2.1.2 Snowball 11](#_Toc480974745)

[2.1.3 KnowItAll 12](#_Toc480974746)

[2.1.4 TextRunner 13](#_Toc480974747)

[2.2 基于有监督学习的关系抽取 15](#_Toc480974748)

[2.2.1 基于特征向量的方法 15](#_Toc480974749)

[2.2.2 基于核函数的方法 16](#_Toc480974750)

[2.3 基于深度学习的关系抽取 17](#_Toc480974751)

[2.3.1 基于CNN的方法 18](#_Toc480974752)

[2.3.2 基于RNN的方法 19](#_Toc480974753)

[2.4 本章小结 20](#_Toc480974754)

[第三章 句子的分布式表示 21](#_Toc480974755)

[3.1 引言 21](#_Toc480974756)

[3.2 词向量加权 24](#_Toc480974757)

[3.3 基于CNN的句子分类算法 26](#_Toc480974758)

[3.3.1 位置嵌入 28](#_Toc480974759)

[3.3.2 卷积和池化 29](#_Toc480974760)

[3.3.3 Dropout和L2正则化 29](#_Toc480974761)

[3.3.4 反向传播训练 30](#_Toc480974762)

[3.4 对比实验 31](#_Toc480974763)

[3.4.1 数据集及评价标准 31](#_Toc480974764)

[3.4.2 实验设置 32](#_Toc480974765)

[3.4.3 实验结果及分析 33](#_Toc480974766)

[3.5 本章小结 34](#_Toc480974767)

[第四章 面向互联网新闻文本的企业关系抽取 36](#_Toc480974768)

[4.1 引言 36](#_Toc480974769)

[4.2 网页正文提取 36](#_Toc480974770)

[4.3 企业实体识别 40](#_Toc480974771)

[4.4 基于Bootstrapping构建语料的方法 42](#_Toc480974772)

[4.4.1 关系类型构建 43](#_Toc480974773)

[4.4.2 初始种子集构建 45](#_Toc480974774)

[4.4.3 句子的相似度计算及聚类 45](#_Toc480974775)

[4.4.4 语料库去重 49](#_Toc480974776)

[4.5 实验 50](#_Toc480974777)

[4.5.1 数据集及评价标准 50](#_Toc480974778)

[4.5.2 实验设置 51](#_Toc480974779)

[4.5.3 实验结果及分析 52](#_Toc480974780)

[4.6 本章小结 53](#_Toc480974781)

[第五章 总结与展望 54](#_Toc480974782)

[5.1 工作总结 54](#_Toc480974783)

[5.2 未来展望 54](#_Toc480974784)

[参考文献 55](#_Toc480974785)

[致谢 58](#_Toc480974786)

[附录 59](#_Toc480974787)

1. 绪论
   1. 研究背景及意义

随着互联网的普及和发展，信息量正以指数规律飞速地增长，每天数以亿计的文本数据在互联网上不断更新，这其中包括新闻、社交、政府网站数据。在这些数据当中蕴含着许多对人们有价值的信息，这些信息对人们的生产生活起着至关重要的作用。然而面对这些海量的互联网数据，单凭人力很难快速地从中获取自己所需要的信息。为了应对信息过载带来的挑战，迫切需要一些自动化的方法帮助人们从中迅速找到真正有用的信息。

信息抽取 (Information Extraction, IE) 研究正是在这种背景下产生的。信息抽取作为自然语言处理(Natural Language Processing，NLP)的一项经典任务主要目的是从自然语言文本中抽取出指定的实体(Entity)、关系(Relation)、事件(Event)等事实信息。信息抽取可以将无结构化的文本信息转化为结构化或半结构化的信息，并以数据库的形式存储，一方面可以用于对文本的快速阅读和理解，帮助人们更方便的获取所需要的信息，另一方面可以用于深入地挖掘分析，对知识库构建、垂直搜索、自动问答等自然语言处理相关领域起着非常重要的作用。其中，实体关系抽取(Entity Relation Extraction)作为信息抽取的子任务和关键技术之一，近年来受到越来越多研究者的关注，逐渐成为研究热点。

实体关系抽取的主要任务是识别出文本中的实体并抽取实体之间的语义关系[1]。例如，句子“Bill Gates is the founder of Microsoft Inc.”中包含一个实体对(Bill Gates, Microsoft Inc.)，以及这两个实体对之间的关系为Founder。通过以上介绍可以发现，如果说信息抽取的主要功能是自动将非结构化的自然语言文本表述为结构化的表格数据，实体识别确定了表格中各个元素的话，那么实体关系抽取则是确定这些元素在表格中的相对位置。由此可见，实体关系抽取的主要目的是在实体识别的基础上，把无结构的自然语言文本中所蕴含的实体之间的语义关系抽取出来，整理成结构化的三元组(关系，实体1，实体2)存储在数据库中，供用户查询或者进一步分析利用。

传统的实体关系抽取的方法一般需要预先定义好实体关系类型体系，针对预先定义好的每一类实体关系人工标注训练数据集，最后基于统计机器学习的方法构建关系的分类模型。然而这种做法的一个主要缺陷是模型的最终分类效果将严重依赖于所提取的特征的质量，而这些特征通常是由预先存在的自然语言处理工具产生的。实际情况是这些处理工具不可避免的会引入一些误差，这些误差将传递到关系分类的模型当中，影响最终的分类效果。

近些年来，深度学习技术(Deep Learning)[2]在图像识别[3]、语音识别[4]和自然语言处理[5]等多个人工智能领域都取得不错的效果。深度学习基于一种深层的非线性网络结构，实现对复杂函数的逼近。深度学习算法和一些浅层的学习算法相比，它会使用更多的层来对输入数据进行处理，因而可以学习到更加丰富的特征。从仿生学角度来说，深度学习的原理与人的大脑皮层一样，分层对输入数据进行处理。在每一层当中，信号被处理单元接受并做转换，如同神经元接受电信号一样。参数通过训练学习得到，最终得到数据的本质特征[6]。相比于传统的机器学习技术，深度学习技术可以自动学习到数据的特征，将人们从繁杂的特征工程中解放出来。

最近，很多研究人员开始将深度学习的技术应用到关系抽取中。借助于词向量(Word Embedding)[[1]](#footnote-1)和深层的神经网络(Deep Neural Network, DNN)[[2]](#footnote-2)，可以学习得到句子级别和文档级别的分布式向量表示。该向量包含了文本的语义信息，可以作为关系抽取分类模型的输入。现有的基于深度学习技术的关系抽取方法都取得了state-of-the-art的效果。基于深度学习技术的关系抽取研究已经逐渐成为相关研究领域的热点。

实体关系抽取在NLP领域具有许多应用，如文档摘要、机器翻译、知识图谱构建、实体消歧、语言建模等。由此可见对实体关系抽取的研究是十分有意义的。

* 1. 关系抽取研究现状

信息抽取目的是从大规模非结构或半结构的自然语言文本中抽取结构化信息。关系抽取是其中的重要子任务之一，主要目的是从文本中识别出实体并抽取实体之间的语义关系。给定一个句子并标注好句子当中的实体对和，关系抽取的任务就是识别出和之间的语义关系。针对关系抽取任务，许多方法已经被提出来，包括基于无监督学习的方法和基于有监督学习的方法。

在基于无监督学习的相关方法中，有的利用了实体所在的上下文信息。它是基于一种叫做分布假说的理论(Distributional hypothesis theory)[7]：出现在同一篇文章中的词往往有着相近的意思。该假说相应的可以扩展为：出现在相似文章当中的实体对往往具有类似的关系。Hasegawa[8]采用层次聚类方法来聚类实体的上下文，并简单地选择了上下文中出现最频繁的词来代表实体之间的关系。Chen[9]提出了一种基于模型顺序选择和标签识别的无监督方法，以解决这个问题。

有监督的学习方法是将关系抽取任务当成多分类问题。根据训练数据设计有效的特征，从而学习各种分类模型，然后使用训练好的分类器预测关系。大体上，这些方法可以分成两类：基于特征向量的方法和基于核函数的方法。在基于特征向量的方法中，通常做法是将句子的一些信息例如词序列、解析树等转换成特征向量[10]。该方法在将结构化表示转换成特征向量时通常面临选择合适的特征集的问题。而基于核函数的方法则无需显示地抽取出特征，但需要定义一个比较好的核函数，目前已有的工作包括：基于卷积树的核函数[11]，子序列核函数[12]以及基于依存语法树的核函数[13]。无论是基于特征向量方法还是基于内核函数的方法，它主要取决于实体在上下文中的各种词汇，句法，语义和其他信息，或背景知识，以提高算法的性能。因此，如何提取和有效地使用更有用的关系提取的词汇、句法、语义等特征，即特征提取和特征选择，已经成为监督关系提取方法研究的重点。

基于有监督学习的关系抽取方法还面临着一个问题，就是缺少足够多的标注数据用于模型的训练，而人工标注训练数据需要花费大量的时间和精力。针对这个局限性，Mintz 等人[14]提出了远程监督(Distant Supervision)的思想。作者们将纽约时报新闻文本与大规模知识图谱 Freebase[[3]](#footnote-3)(包含7300多个关系和超过9亿的实体)进行实体对齐。远程监督基于这样一种假设，一个同时包含两个实体的句子蕴含了该实体对在 Freebase 中的关系，并将该句子作为该实体对所对应关系的训练正例。作者在远程监督标注的数据上提取文本特征并训练关系分类模型，有效解决了关系抽取的标注数据规模问题。但是该方法存在数据噪声问题，即标注的数据中可能存在错误。为了解决这个问题，之后许多研究者从各个角度对远程监督技术提出了改进方案。例如Takamatsu等人[15]改进了实体对齐的技术，降低了数据噪音，提高了关系抽取的总体效果。Yao 等人[16]提出了基于无向图模型的关系抽取方法。Riedel 等人[17]则增强了远程监督的假设，与 Mintz相比错误率减少了 31%。

以上远程监督技术都假设一个实体对只对应一种关系。但是很多实体之间具有多种关系。例如，“Steve Jobs”和“Apple”之间存在多种关系：“Steve Jobs founded Apple”和“Steve Jobs is the CEO of Apple”。因此，Hoffmann等人[18]提出采用多实例多标签(Multi-Instance Multi-label)方法来对关系抽取进行建模，刻画一个实体对可能存在多种关系的情况。类似地，Surdeanu等人[19]也提出利用多实例多标签和贝叶斯网络来进行关系抽取。

现有的有监督学习关系抽取方法已经取得了较好的效果，但它们严重依赖词性标注、句法解析等自然语言处理标注提供分类特征。而自然语言处理标注工具往往存在大量错误，这些错误将会在关系抽取系统中不断传播放大，最终影响关系抽取的效果。最近，很多研究人员开始将深度学习的技术应用到关系抽取中。研究人员使用DNN(Deep Neural Network)来自动学习特征。Socher[20]提出使用递归神经网络来解决关系抽取问题，该方法首先对句子进行句法解析，然后为句法树上的每个节点学习向量表示。通过递归神经网络，可以从句 法树最低端的词向量开始，按照句子的句法结构迭代合并，最终得到该句子的向量表示，并用于关系分类。该方法能够有效地考虑句子的句法结构信息，但同时该方法无法很好地考虑两个实体在句子中的位置和语义信息。

* 1. 本文主要工作

本文主要工作围绕实体关系抽取任务的相关问题展开，主要包括两个部分，分别是句子的分布式表示和面向互联网新闻当中企业实体关系的抽取。需要注意的是，本文只考虑句子级别上的关系抽取。

第一部分针对传统词袋模型在表征句子时缺乏语义信息以及未考虑词的位置信息的缺陷，提出了基于词向量加权和基于卷积神经网络的方法用于构建紧凑且具有语义的句子分布式表示，作为构建关系分类模型的输入。

第二部分针对互联网新闻当中存在的企业实体关系，提出了基于Bootstrapping技术构建关系语料库的方法，克服了纯人工标注过程中费时费力的缺点。并在第一部分的工作基础上，结合网页正文提取、命名实体识别等关键技术，构建了面向互联网新闻文本的企业关系抽取系统的原型。

* 1. 本文组织结构

本文后续组织结构如下：

第二章介绍了当今关系抽取领域的相关研究和基本方法，包括基于有监督学习的方法、基于半监督学习的方法和基于深度学习的方法等，其中较为详细地介绍了基于深度学习的方法，因为它们和本文的工作比较接近。并在最后列举出了一些关系抽取的评价策略。

第三章设计了基于词向量加权的方法和基于卷积神经网络的方法用于构建句子的分布式表示，作为关系分类模型的输入。并在标准数据集SemEval-2010 Task 8上做了一组对比实验，实验结果表明两种方法都能得到紧凑且具有语义的句子的分布式表示。

第四章主要是在前一章工作的基础上，介绍了面向互联网新闻中企业实体关系抽取的一些关键技术和实现过程。其中包括基于Bootstrapping技术构建关系语料库的方法、爬虫模块设计与实现、网页正文提取技术、命名实体识别技术等。

第五章是总结与展望,包括对论文系统的主要工作进行总结并针对本文工作的可完善之处进行展望和探讨。

1. 关系抽取的相关研究
   1. 基于半监督学习的关系抽取

对于很多自然语言处理任务来说，包括关系抽取，常常有很多现成的未标注数据，然而要获取大量标注数据则需要花费相当多的时间和精力，因此一些基于bootstrapping的半监督学习技术越来越受到研究者的关注。半监督学习技术的目标就是利用这些大量的、廉价的无标记数据帮助我们得到更好的训练模型。

本文将注意力集中在Yarowsky[21]和Blum＆Mitchell[22]的算法上，因为之后大部分基于半监督学习的关系抽取方法都使用了这种算法（Co-training[[4]](#footnote-4)）。它们的主要思想是使用弱学习器的输出作为下一次迭代的训练数据(输入)。协同训练是一种弱监督模型，它从一小组标记数据以及使用分开但冗余的大量未标记数据（即使用不相交的特征子集来表示数据）中进行学习任务。为了确保模型的有效性，协同训练（Co-training）要求从两个不同的视角看待训练数据，它假定每个样例使用两个不同的特征集来描述，这两个特征集提供关于样例不同的补充信息。同时特征集必须满足两个条件：(1)特征集对于目标学习任务来说是有效的，(2)对于不同的特征集需要使用独立的分类器去训练。

|  |
| --- |
| 算法2.1：Yarowsky’s algorithm in general form |
| **输入：**一组未标记的数据集*D*以及一组种子数据集*S*  **重复**  在数据集S上训练出分类器C  用分类器C去标记数据集D  *N* = top *n* labels that *C* was highly confident  *S = S*∪*N*  *D = D*\*N*  **直到** 达到收敛标准 |

Yarowsky在论文中描述的算法框架已经在算法2.1给出[23]。Yarowsky将这个算法应用到了词义消歧任务，并取得了不错的效果。他的这个算法本质上是协同训练算法的一个特例。Abney[24]提出了Yarowsky算法的理论分析。当探究相关协同训练的算法时，需要考虑两个主要问题：1)如何自动地获取种子数据集2)什么样的种子集是好的，即如何评价种子集。目前效果比较好的基于半监督的关系抽取系统有：DIPRE[25]、Snowball[26]、KnowItAll[27]以及TextRunner[28]。

* + 1. DIPRE

DIPRE(Dual Iterative Pattern Relation Expansion)是由Brin在1998年设计出的关系抽取系统。它主要关注的关系是互联网上(作者，书目)这样一种关系。DIPRE从一小组集合对(作者，书目)开始，也就是初始种子集。现在假设初始种子集中只有一个条目(*Arthur Conan Doyle, The Adventures of Sherlock Holmes*).，接着该系统会去互联网上爬取所有包含种子条目的网页(作者、书目必须同时存在该网页中)。为了归纳出其中的关系模式，DIPRE使用一个六元组来表示这一模式[顺序，作者，书名，前缀，后缀，中间]。其中顺序是1的话表示作者在前书名在后，0则代表相反，前缀后缀则表示两个实体的头部和尾部所出现的字符，中间则代表两个实体之间所出现的字符。

举个例子，假设从互联网上爬取到如下文本数据：

*“Read The Adventures of Sherlock Holmes by Arthur Conan Doyle online or in you email”*

*“know that Sir Arthur Conan Doyle wrote The Adventures of Sherlock Holmes, in 1892”*

*“When Sir Arthur Conan Doyle wrote the adventures of Sherlock Holmes in 1892 he was high …”*

接着抽取句子当中的六元组：

[*0, Arthur Conan Doyle, The Adventures of Sherlock Holmes, Read, online or, by*]

[*1, Arthur Conan Doyle, The Adventures of Sherlock Holmes, know that Sir, in 1892, wrote*]

[*1, Arthur Conan Doyle, The Adventures of Sherlock Holmes, When Sir, in 1892 he, wrote*]

在抽取完所有句子的六元组之后，系统会将那些顺序值相同以及中间字符串相同的元组分到同一组当中。针对每一组，接下来系统将找出它们最长公共前缀和最长公共后缀，最终将归纳出每一组的关系模式：

[*最长公共前缀, 作者, 中间部分, 书名, 最长公共后缀*].

对于上面的例子，得到的关系模式将会是[*Sir, Arthur Conan Doyle, wrote, The Adventures of Sherlock Holmes, in 1892*]，下一步是使用通配符表达式来泛化抽取模式，于是将得到：[*Sir, .\*?, wrote, .\*?, in 1892*]。然后使用这样一个模式再去互联网上搜索，并假设得到一个新的关系：(Arthur Conan Doyle, The Speckled Band)。DIPRE会将它加入到种子集当中并重复上述过程直到满足停止条件。算法2.2[29]描述了整个系统处理的过程。

|  |
| --- |
| 算法2.2：DIPRE |
| 1. 使用初始种子集去互联上标记一些数据 2. 从标记的样例数据中归纳出关系模式 3. 将上一步归纳出的关系模式去互联网上进行匹配搜索，并得到新的关系对，接着添加到种子集当中 4. 重复上述步骤直到满足停止条件 |

DIPRE和Yarowsky算法相似之处在于：算法的初始化都是从一小组初始种子集开始的，DIPRE使用的分类器就是一个模式匹配，这个关系抽取模式是从种子关系集开始迭代训练得到的。给定一个字符串，如果它和某种模式相匹配，那么它将被归为正类，并用于抽取新的关系样例。并且新的样例会被加入到种子集当中，迭代训练后会得到更多的关系模式和样例。DIPRE可以看成是Yarowsky算法应用于关系抽取的范例。

通过DIPRE的算法也可以发现，它的主要缺点是模式匹配系统比较严格。例如两个模式只是单个标点符号不同，那么它们在DIPRE中就是不同的。这样一种匹配方式将会大大降低系统的召回率。

* + 1. Snowball

Snowball的系统结构和上面所提到的DIPRE有点类似，它的目标是识别文本当中的(公司，位置)这样一种关系。Snowball也会从一个初始种子关系集开始，它所使用的分类器和DIPRE一样是一个匹配模式*P*，但是这当中又有区别，Snowball会将每个元组表示成一个向量，然后用一个相似性函数来将它们分组，而不是直接去匹配字符串。Snowball会从原始文本中抽取出如下形式的元组[*prefix, orgnization, middle, location, suffix*]，例如(*CMU, Pittsburgh*)是种子关系集合中的一个实体关系对，对于句子：“*... go to CMU campus in Pittsburgh to meet ...*”系统将会提取出：

[*(w1, go), (w2, to), ORG, (w1, campus), (w2, in), LOC, (w1, to), (w2, meet)*]

其中*prefix*和*suffix*特征向量的长度限制在了2个单词。代表某个词的权重，它是通过计算该词归一化的频率得到的。例如对于出现在*suffix*位置的单词*meet*权重计算如下：

|  |  |
| --- | --- |
|  | (2-5) |

随着种子集的扩充，更多的元组将被加入进来，与此同时单词的权重也在不断的更新和调整。同时Snowball也给出了元组之间的相似度计算公式：

有了相似度计算的方法，就可以对相似性较高的元组进行聚类，对于每一类元组，Snowball会推导出一个中心向量作为该类元组的模式P的一个表示。并且每一类模式P都会被赋予一个置信度的评分用于评价该模式的质量：

|  |  |
| --- | --- |
|  | (2-6) |

其中表示在之前一轮的迭代训练过程所有满足关系模式P并且（organization, location）完全一样的样本数量，则表示那些满足关系模式P但（organization, location）不完全相同的样本数量。

为了标记新的数据样本，Snowball会在原始数据上运行一遍命名实体识别程序来识别出文本当中所有的地理位置和公司实体。在一个句子当中，对于每一个(organization, location)关系对系统都会产生一个五元组，因此一些出现频率较高的关系对会有很多个元组和它相对应，系统将每个候选关系元组和所有模式进行匹配，最终仅保留那些相似性得分大于某个阈值的候选元组。接下来再根据所匹配模式*P*的置信度来给候选关系对评分。最终那些评分较高的会被加入到种子集合当中并进行下一轮的迭代。

与DIPRE相比，Snowball的匹配系统更加的具有灵活性。Snowball不仅仅停留在文本表面的精确匹配，而允许文本有些微小的变化，例如标点符号等。

* + 1. KnowItAll

与DIPRE和Snowball不同，KnowItAll是一个大规模的Web信息抽取系统，它只使用一小组领域无关的提取模式来标记自己的训练样例。当为特定关系实例化时，这些通用模式产生特定关系的提取规则，然后用于学习特定领域的提取规则。这些抽取规则可以被应用于互联网网页当中，例如识别搜索引擎的查询语句，并且使用从搜索引擎命中计数导出的逐点互信息（PMI）给所得到的抽取结果分配一个置信度。例如，KnowItAll利用诸如“<NP1>如<NP2>”的通用提取模式来建议NP2作为类NP1的候选成员的实例化。接下来，系统使用频率信息来识别哪些实例化最有可能是类的成员。最后，KnowItAll学习一组关系特定的提取模式，例如“国家的首都”，从而它可以提取更多的城市。

* + 1. TextRunner

DIPRE，Snowball和KnowItAll都是关系特定的系统，即已经指定了特定种类的关系。系统所涉及的关系集合必须事先由人工定义好，而TextRunner的提出正是为了克服这一缺点。TextRunner 是第一个开放式实体关系抽取系统，不需要人工定义关系类型体系，它以自我监督的方式从其语料库中的文本中学习关系，种类和实体。TextRunner以三元组的形式定义一个关系，其中和是旨在表示实体或名词短语的字符串，*r*是表示和之间的关系的字符串。图2-1是TextRunner的系统框架图。



图2-1TextRunner系统框架图

TextRunner包含三个主要的模块，分别是：1)训练分类器模块；2)大规模Web文本抽取模块；3)可信度计算模块。下面将详细介绍这三个模块。

1. 训练分类器模块

TextRunner使用启发式规则自动从小规模标注语料中构建训练语料，然后再训练贝叶斯分类器。

1. 利用启发式规则自动构建语料(基于依存句法分析)，其规则如下：a)的词性标注序列；b)的长度；c)包含停用词的数目；d)实体是否是专有名词；e)左边词语的词性；f)右边词语的词性。
2. 训练贝叶斯分类器，对每一个三元组()进行分类，其特征如下：a)的词性标注序列；b)的长度；c)包含停用词的数目；d)实体是否是专有名词；e)左边词语的词性；f)右边词语的词性。
3. 大规模Web文本抽取模块

对大规模Web文本中的每个句子当中的三元组使用贝叶斯分类器进行分类，筛选出被标记为正例的三元组并存储。

1. 可信度计算模块

合并相似的三元组，如()和( )是两个相同的三元组，计算合并后的三元组在Web文本中出现的句子数目，把该句子数目作为三元组的一个可信度。

TextRunner最大的问题在于系统严重依赖依存句法分析工具来实现训练数据的自我标注，假如数据标注质量不高，这将直接影响接下来的抽取效果。

表2-1 几种基于半监督的关系抽取系统对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **DIPRE** | **Snowball** | **KnowItAll** | **TextRunner** |
| Initial seed | *Yes* | *Yes* | *Yes* | *No* |
| Predefined relation | *Yes* | *Yes* | *Yes* | *No* |
| External NLP tools | *No* | *Yes:NER* | *Yes:NP chunker* | *Yes:dependency parser, NP chunker* |
| Relation types | *Binary* | *Binary* | *Unary / Binary* | *Binary* |
| Language dependent | *No* | *Yes* | *Yes* | *Yes* |
| Classifier | *Exact pattern matching* | *Matching with similarity function* | *Naïve Bayes classifier* | *Self-supervised binary classifier* |
| Input parameters | *2* | *9* | *>=4* | *N/A* |

表格2-1给出几种基于半监督关系抽取系统的对比，主要从是否需要初始种子集、是否需要预先定义关系集、是否需要外部的NLP预处理工具、关系类型、所使用的分类器以及输入参数数量这几个维度考察了这几类系统。Snowball, KnowItAll和TextRunner这几个系统初始化时需要输入一定数量的参数，相关参数的定义在系统中都有明确解释，并提供了一些选择给用户。然而，这几个系统都没有给出参数调优的方法，用户不知道如何初始化参数以达到系统的最优效果，只能通过经验来选取。

* 1. 基于有监督学习的关系抽取

有监督的关系抽取系统通常需要大量人工标注的训练数据，从训练数据中自动学习关系对应的抽取模式[30]。有监督关系抽取方法主要包括：基于特征向量的方法[31]，基于核函数的方法[32]，基于句法解析增强的方法[33]和基于条件随机场的方法[34]。

对于给定的句子,其中和代表实体，关系映射函数可以定义如下：

|  |  |
| --- | --- |
|  | (2-1) |

公式(2.1)中T(S)是从句子S中提取的各类特征，映射函数用来决定句子中的两个实体是否包含关系R。换句话说，关系抽取任务转变成为了*实体-关系的发现*。当我们拥有一批正负关系样例的训练数据后，我们可以将构建成为一个分类器，例如支持向量机(SVM)，贝叶斯分类器，感知机等。这些分类器可以使用在执行文本分析之后选择的一组特征（如POS标注，依存句法分析等）。另一方面，输入到分类器的特征也可以是丰富的结构表示形式例如解析树。根据输入到分类器特征的性质，正如上文刚刚所提到的，基于有监督的关系抽取方法可以进一步分为：(1) 基于特征向量的方法和(2) 基于核函数的方法。接下来我会适当详细地介绍这两种方法。

* + 1. 基于特征向量的方法

当拥有带有标注的训练数据后，接下来可以从句子当中提取出语法和语义的特征。这些特征可以作为线索帮助判断句子当中的两个实体之间是否存在某种关系。其中能够抽取的语法特征包括：1)实体本身；2)两个实体所属的类；3)两个实体之间的词序列；4)两个实体之间词的个数；5)包含两个实体解析树的路径。语义特征包括两个实体之间的依存句法树路径。无论提取的是语义特征还是句法特征都以特征向量的形式呈现给分类器，用于训练或分类。Kambhatla使用所述的特征训练对数线性模型，用于实体分类任务。另一方面，Zhao和Grishman使用了SVM分类器去训练那些基于多项式和线性核函数的特征，用于分类不同的实体关系。有些特征对于实体关系分类的效果很好，而有的特征则显得较弱，如何筛选出那些与分类任务较为相关的特征则显得十分重要。基于特征方法涉及特征的启发式选择，并且必须在试错法的基础上选择特征以最大限度地提高分类器的准确率。由于一般的NLP应用和关系提取任务需要对输入数据进行结构化表示，因此可能难以达到最优相关特征的子集。为了解决选择合适的特征集的问题，人们设计了一些专门的核函数用于关系提取，以便挖掘出输入数据的丰富表示例如浅层解析树等。这些核函数以更隐蔽的方式在更高维度的空间中穷举地探索输入表示。接下来将详细介绍相关基于核函数的方法。

* + 1. 基于核函数的方法

一般用于关系抽取的核函数是基于string-kernels[[5]](#footnote-5)的，string-kernels过去一直被应用于文本分类的应用当中[35]，它是一个操作字符串的核函数，用于度量两个字符串之间的相似性。如果字符串a和b越相似，那么string-kernels函数的值就越大。同时每个字符串都可以映射到更高维的空间，其中每个维度对应于特定子序列的存在（加权）或不存在（表示为0）。例如，字符串可以在子序列的更高维空间表示如下：

|  |  |
| --- | --- |
|  |  |
|  | (2-2) |

其中 表示衰减因子，使得较长的和不连续的子序列受到惩罚。不妨设表示字符串的一个子串，的下标，表示子串的长度。由于可以以多种方式存在于内部，因此字符串的坐标对应到在高维空间中的位置可以表示为

|  |  |
| --- | --- |
|  | (2-3) |

如果*U*是存在于字符串*x*和*y*中的所有可能有序子序列的集合，则*x*和*y*之间的相似度可由下式给出：

|  |  |
| --- | --- |
|  |  |
|  | (2-4) |

直接使用公式(2-3)来计算(2-4)将会是指数级的时间复杂度，因此在实际计算过程当中一般会使用动态规划算法来进行求解。这样时间复杂度会降低至。 对于公式(2-4)一种更为概括的解释是：假设代表两个不同的object，这里的object可以是字符串、词的序列以及句子的解析树等等。在实体关系抽取任务中，如果和分别代表实体关系样本中的正负样例，并且y代表测试样例，则表示包含了这样一种关系反之则不包含。在实际应用当中，一般用某种分类器来表示相似度计算函数，例如SVM，贝叶斯等。、和可以有如下两种表现形式：(1)实体周围的词序列，(2)包含实体的语法分析树。根据表现形式的不同，目前主要有两种基于核函数的实现方法，分别是：Bag of features kernels和Tree kernels。

通过以上介绍可以发现，基于有监督学习的方法存在着以下几点不足：

1. 这些方法难以扩展到新的实体关系类型，因为需要特定领域带有标注信息的关系语料库；
2. 对高阶实体关系(Higher-order Relations)又称多元实体关系的扩展也很困难；
3. 大多数方法都没有考虑对计算复杂度的优化，随着输入数据量的增加，计算可能成为一个瓶颈；
4. 需要利用现存的NLP工具对输入数据进行一些预处理，例如构建句子的解析树、依存语法分析树等。实际情况是这些处理工具不可避免的会引入一些误差，这些误差将传递到关系分类的模型当中，影响最终的分类效果。
   1. 基于深度学习的关系抽取

上述工作已经证明有监督和半监督的方法能够很好地解决关系关系抽取任务，并产生相对较高的性能。然而，这种方法的实际效果严重依赖于所选特征的质量。随着近来深度神经网络的兴起，很多研究人员开始将深度学习的技术应用到关系抽取中。主要有基于卷积神经网络的方法和基于循环神经网络的方法。

* + 1. 基于CNN的方法

卷积神经网络(Convolutional Neural Networks, CNN)现在非常适用于处理自然语言处理的一些任务并且可以得到效果不错的模型。CNN是具有多层的前馈人工神经网络的扩展架构。CNN既可以是基于有监督的，也可以是基于无监督的，但从效果来看，有监督的方法准确率通常会比无监督的方法高一点。人工神经网络基本上都会有输入层，隐藏层和输出层。每个隐藏层和输出层节点连接都模仿动物的视觉皮层的行为，并称作神经元。而CNN会在输入层上应用卷积来计算输出。因此，局部连接会被创建，其中输入的每个区域连接到输出神经元。CNN专为最小化预处理而设计，研究人员针对不同问题提出了不同架构的CNN模型，以提高关系分类性能。

理解卷积的最简单的方法是将滑动窗函数应用于矩阵。滑动窗口可以命名为过滤器，特征检测器或内核。为了实现完整的卷积滤波器，将其值与原始矩阵进行乘积，然后通过在整个矩阵上滑动滤波器来为每个元素求和。CNN中的第一层是输入层。 它可以具有单个或多个通道，取决于表示和需要，或者可以用于不同词向量的单独通道。第二层或卷积层由特征图谱组成。要从输入层移动到特征图，输入层与滤波器进行卷积，然后在池化层添加偏置。池化层仔细检查其输入，并在每个过滤器的结果处取最大值是一个常见方法。最终结果送入一个非线性层，即一个带有激活功能的神经元层。滤波器会被随机初始化，并在模型训练中被迭代更新。不同的滤波器互不相同，并且每个滤波器都对应着一个特征图谱。

最后，在经过多个卷积层和池化层之后，最后会有一个全连接层与之相连接。全连接层的神经元会与先前层中的所有激活的神经元完全连接，并且它们能否被激活可以通过矩阵乘法来评估。 通过max pooling生成的特征向量会被送到损失层。损失层使用损失函数作为网络训练的惩罚并产生预测和真实标签的变化。从而输出层可以提取出输入句的关系标签。研究人员会选择超参数和正则化方案来解决其模型中的过拟合问题，以增加模型的泛化能力。

Zeng[36]提出采用卷积神经网络进行关系抽取。他们采用词汇向量和词的位置向量作为卷积神经网络的输入，通过卷积层、池化层和非线性层得到句子表示。通过考虑实体的位置向量和其他相关的词汇特征，句子中的实体信息能够被较好地考虑到关系抽取中。Zeng的主要贡献可以总结为三个方面：1) 探索在没有复杂NLP预处理的情况下执行关系分类的可行性，其中卷积神经网络主要用于提取词法和句子级特征；2) 为了指定应该分配关系标签的实体对，提出了位置特征来对CNN中的目标名词对的相对距离进行编码。3) 使用SemEval-2010 Task 8数据集进行实验，实验结果表明，提出的位置特征对关系分类至关重要。提取的词汇和句子级别特征对于关系分类是有效的。后来，Santos[37]还提出了一种新的卷积神经网络进行关系抽取，其中采用了新的损失函数，能够有效地提高不同关系类别之间的区分性。

* + 1. 基于RNN的方法

循环神经网络(Recurrent Neural Network,RNN)是一类人造神经网络，其中单元之间的连接形成定向循环。这创建了网络的内部状态，允许其呈现动态时间行为。与前馈神经网络不同，RNN可以使用其内部存储器来处理任意输入序列。这使得它们适用于诸如未分段连接的手写识别或语音识别等任务。由于基于CNN的方法缺乏学习时间特征的能力，特别是实体对之间的长距离依赖关系处理的效果不够好，因此相关领域的研究学者尝试将RNN应用于关系抽取的任务中。

RNN的思想是利用序列信息，在传统的神经网络中，我们假设所有的输入（和输出）是相互独立的。但是对于许多任务来说事实情况并非如此。比如说预测一个句子中的下一个单词，需要考虑哪些单词来自它。RNN称为循环，因为它们对序列的每个元素执行相同的任务，输出取决于先前的计算。考虑RNN的另一种方法是，它们有一个“记忆”，捕获到目前为止所计算的信息。在理论上，RNN可以以任意长的序列使用信息，但在实践中，它们仅限于回顾前几个步骤。一个典型RNN结构如图2-2所示：



图 2-2 RNN结构示意图

其中是在时间处的输入，举个例子，可以表示句子当中第二个单词对应的One-hot向量表示。代表时间t处的隐藏状态，它是网络的记忆单元。代表时间t处的输出。

Socher[20]提出使用递归神经网络来解决关系抽取问题。该方法首先对句子进行句法解析，然后为句法树上的每个节点学习向量表示。通过递归神经网络，可以从句法树最低端的词向量开始，按照句子的句法结构迭代合并，最终得到该句子的向量表示，并用于关系分类。该方法能够有效地考虑句子的句法结构信息，但同时该方法无法很好地考虑两个实体在句子中的位置和语义信息。

Miwa[38]提出了一种基于端到端神经网络的关系抽取模型。该模型使用双向LSTM(Long-Short Term Memory，长短时记忆模型)和树形LSTM同时对实体和句子进行建模。目前，基于卷积神经网络的方法在关系抽取的标准数据集 SemEval-2010 Task 8上取得了最好的效果。

上面介绍的神经网络模型在人工标注的数据集上取得了巨大成功。然而，与之前基于特征的关系抽取系统类似，神经网络关系抽取模型也面临着人工标注数据较少的问题。对此，Zeng尝试将基于卷积神经网络的关系抽取模型扩展到远程监督数据上。Zeng 假设每个实体对的所有句子中至少存在一个句子反映该实体对的关系，提出了一种新的学习框架：以实体对为单位， 对于每个实体对只考虑最能反映其关系的那个句子。该方法在一定程度上解决了神经网络关系抽取模型在远程监督数据上的应用，在 NYT10数据集上取得了远远高于基于特征的关系抽取模型的预测效果。但是，该方法仍然存在一定的缺陷：该模型对于每个实体对只能选用一个句子进行学习和预测，损失了来自其他大量的有效句子的信息。

* 1. 本章小结

本章主要介绍了实体关系抽取任务现有的主要研究方法，包括：基于半监督学习的方法、基于有监督学习的方法以及基于深度学习的方法。

半监督学习方法主要基于种子集的Bootstrapping方法。该方法首先需要根据预定义的关系类型构建相应的关系实例作为初始种子集，然后通过模式学习方法迭代生成关系描述模式集。半监督学习方法不需要手动注释语料库，只需要构建初始关系种子集，然后使用Web或大规模语料库信息进行高度冗余，充分挖掘相应的关系描述模式，并通过模式匹配提取新关系示例，准确高效地完成关系提取任务。然而，该方法存在几个关键问题，如初始关系种子集的生成和选择，模式模式，模式的质量，迭代过程的速度以及高精度，低召回率等。

监督学习方法主要包括两类：基于特征向量的学习方法和基于内核函数的学习方法。这种方法将关系提取任务作为分类问题。首先需要手动标注大规模的训练语料库，然后在标记语料库的基础上提取和选择特征。通过使用不同的机器学习算法来训练学习分类模型，提取新的实体对。对于监督学习方法，无论是基于特征向量学习算法，还是基于内核函数学习算法，主要依赖于实体在各种词汇，句法，语义等信息或背景知识的上下文中，改进算法的性能因此，如何挖掘和有效地使用更多的词法，句法和语义特征，如何进行有效的特征提取和特征选择已经成为关系提取方法研究的重点。

现有的监督学习关系的方法已经取得了很好的效果，但是它们依赖于一些自然语言处理的工具，如词性标注，句法解析等。而自然语言处理工具往往是大量的错误，这些错误会不断传播，最终影响关系提取的效果。因此许多研究人员开始将深度学习的技术应用于关系抽取。基于卷积神经网络的方法被用关系抽取的输入，通过使用词汇向量和词的位置向量，句子通过卷积层，池层和非线性层得到特征表示。通过考虑实体的位置向量和其他相关的词的特征，可以在关系提取中更好地考虑句子中的实体信息。

1. 句子的分布式表示
   1. 引言

在实现关系抽取的相关任务时，通常做法是将其转换为句子的分类任务，即事先定义好几种关系，然后构建关系的分类模型。而在处理NLP相关问题时，除了诸如分词、去停用词、词根化(Stemming)等一系列数据清洗的必要步骤外，第一步肯定是要找一种方法将文本中的字词转为向量形式表达(Vector Representations)。 自然语言处理当中最直观，最常用的词表示方法是One-hot Representation[[6]](#footnote-6)，它是一种局部表示(Local Representation)[42]方法，这种表示方法把每个词表示为一个很长的向量。该向量的维度一般是词表大小，并且其中某一个维度的值为1外，其它元素都为0，这个为1的维度就表示当前词。举个例子:

“中国”表示为 [0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 …]

“北京”表示为 [0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 …]

每个词都是茫茫 0 海中的一个 1。

这种 One-hot Representation 如果给每个词分配一个数字 ID，将会是非常的简洁的。比如上述例子中，“中国”记为 3，“北京”记为 8(假设从 0 开始记)。在实际编程过程当中，通过Hash函数给每个词分配一个编号就可以了。然而这种表示方法存在一个问题，即对特征的编码往往是随机的，没有提供任何关联信息，没有考虑到字词间可能存在的关系，也就是“词汇鸿沟”现象。例如，我们对“中国”和“北京”的从属关系，地理位置关系等一无所知，我们从3和8这两个值看不出任何信息。此外，这种表示方法还容易发生维数灾难，尤其是在深度学习相关的一些应用中。

使用分布式表示(Distributed Representations)则可以有效地解决这个问题。分布式表示模型可以将字词转为连续值(相对于One-Hot编码的离散值)的向量表达，并且其中意思相近的词将被映射到向量空间相近的位置。分布式表示模型在NLP中主要依赖的假设是Distributional Hypothesis[43],即在相同的语境中出现的词其语义也相近。Word2Vec[44]即是一种计算非常高效的，可以从原始语料中学习字词空间向量的预测模型，它既能表示词本身又可以考虑了词与词之间的语义距离。它主要分为CBOW(Continues Bag of Words)和Skip-Gram两种模式，其中CBOW是从原始语句(比如:中国的首都是\_\_)推测目标字词(比如:北京)；而Skip-Gram则正好相反，它是从目标字词推测出原始语句，其中CBOW对小型数据比较合适，而Skip-Gram在大型语料中则表现更好。在本节主要使用Skip-Gram模式的Word2Vec。

对于句子和文本的分类任务来说，只有字词的语义空间向量表示还是不够的，还需要得到句子和文本的空间向量表示。最经典的文本向量表示方法要数词袋模型(Bag of Words,BOW)[[7]](#footnote-7)了。通过词袋模型得到的文本向量可以直接输入到分类器，进行文本分类或情感分析等任务。下面通过一个例子来说明词袋模型的建立的过程。

下面是两个简单的例句：

(1) John likes to watch movies. Mary likes movies too.

(2) John also likes to watch football games.

根据这两个样本建立一个词典包含所有出现的单词，如果是未清洗过的原始文档，则以词根建立。

[ "John", "likes", "to", "watch", "movies", "also", "football", "games", "Mary", "too" ]

根据每个词语在词典的索引和在文档中的出现频次可以对以上两个句子建立词袋向量：

(1) [1, 2, 1, 1, 2, 0, 0, 0, 1, 1]

(2) [1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

每个维度的权值可以取词频或是tf\*idf，甚至是二元值，即1代表出现0代表未出现。在得到相关文本的向量表示后就可以进行其它一些比较复杂的NLP任务了。

然而词袋模型也存在两个比较明显的缺陷：一是它丢失了词与词之间的顺序信息，二是它没有考虑词的语义信息。因此词袋模型在NLP有些任务中的表现难免差强人意。

另一种向量化模型和单词输入顺序有关，可以区分"Mary loves Jack"和"Jack loves Mary"这两个句子。词语的顺序很难直接量化输入，但循环神经网络(Recurrent Neural Network, RNN)能够通过时间序列的变化，实现变长词串到语义向量的映射(sequence-vector)。

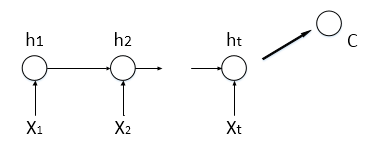


图3.1 RNN结构示图

如图，为输入文档的词串，RNN将文档看成一个随时间变化的词序列，每输入一个新词，隐含层就进行一次更新:

|  |  |
| --- | --- |
|  | (3-1) |

这样，隐含层充分利用了上文的历史信息，并始终保持最新状态，直到最后输出文档的语义向量*c*。由于带有词序信息，RNN训练出来的文本向量相比词袋模型更具有语义，在实验上直接体现在输入到同个分类器中，带词序信息的模型在情感分类任务的表现更好。此外，实现了变长串到向量(sequence-vector)映射的模型，可以适用更复杂的NLP任务[45]，比如配合语言模型可以很容易扩展成sequence-vector-sequence模型，以完成直接从源语言到目标语言的机器翻译任务[46]。然而RNN优化困难、结构复杂，且容易丢失较久之前的历史信息。

本文中尝试使用了两种方法学习句子的紧凑的分布式表示，具体过程将在3.2和3.3小节阐述。

* 1. 词向量加权

词向量(Word Embedding)[[8]](#footnote-8)是自然语言处理中语言模型与表征学习技术的统称。概念上而言，它是指把一个维数为所有词的数量的高维空间嵌入到一个维数低得多的连续向量空间中，每个单词或词组被映射为实数域上的向量。训练词嵌入的方法包括人工神经网络[47]、对词语同现矩阵降维[48][49][50]、概率模型[52]以及单词所在上下文的显式表示等[53]。

word2Vec[[9]](#footnote-9)是由谷歌开源的一套基于神经网络的词嵌入的高效训练工具。Word2Vec是一个双层神经网络，它的输入是文本语料，输出则是一组向量：该语料中词语的特征向量。使用Word2Vec训练语料能得到一些非常有趣的结果，比如意思相近的词在向量空间中的位置会接近。从一份Google训练超大语料得到的结果中看，诸如Beijing、London、New York等城市的名字会在向量空间中聚集在一起，而Cat、Dog、Fish、等动物词汇也会聚集在一起。同时，如图3.2所示，Word2Vec还能学会一些高阶的语言概念，比如计算“man”到“women”的向量(单词都是向量空间中的点，可计算两点之间的距离)，会发现它和“king”到“queen”的向量非常相似，即模型学到了男人与女人的关系；同时，“walking”到“walked”的向量和“swimming”到“swam”的向量非常相似，模型学习到了进行时与过去时的关系。

预测模型Neural probabilistic language models通常使用极大似然法进行训练，在给定前面的语句*h*的情况下，通过一个softmax函数来最大化目标词汇的概率。

|  |  |
| --- | --- |
|  |  |
|  | (3-2) |

其中，用于计算词汇和上文的兼容性(通常使用向量的点积来计算)。一般通过最大化训练集上的对数似然函数来训练这个模型，例如：

|  |  |
| --- | --- |
|  |  |
|  | (3-3) |

但这个方法存在一个比较严重的问题是计算的开销太大，因为在每一步训练过程当中需要计算并且正则化当前上下文环境*h*中所有其它单词的概率得分。在Word2Vec的CBOW模型中，不需要计算完整的概率模型，只需要训练一个二元分类模型，用来区分真实的目标词汇和假想的词汇(噪声)这两类。这种用少量噪声来估计的方法，类似于蒙特卡洛模拟[53]。从数学角度来看，我们的目标是最大化:

|  |  |
| --- | --- |
|  | (3-4) |

其中是一个概率值，是通过目标单词使用二分类逻辑回归计算得出的。在实际中，我们通过从噪声分布中绘制k个对比词来近似期望值。当真实的目标单词被分配到较高的概率，同时噪声单词被分配的概率很低时，目标函数也就达到最大值。

在得到每个词语的词向量表示后，对句子当中每个词进行加权取平均后即可得到句子的语义向量表示，用数学公式表示为：

|  |  |
| --- | --- |
|  | (3-5) |

其中是该文档中出现过的所有词，表示该词的词向量，为词嵌入的加权系数函数,为句子当中词语的总数。

词向量的权重函数*weight()*要能够反映词在所在文段中的重要性。单词的重要性主要体现在两个方面：统计信息和位置信息。统计信息可以用tf\*idf值衡量，位置信息可以通过该词是否在标题出现过来判断：

|  |  |
| --- | --- |
|  | (3-6) |

其中是位置信息函数，如果在标题出现值为1，反之为0；是统计信息函数，为的tf\*idf值。是个增益参数，用来放大在标题中出现过的词的权重，本文取值0.25。

* 1. 基于CNN的句子分类算法

在机器学习中，卷积神经网络(Convolutional Neural Network, CNN)[[10]](#footnote-10)是受生物学上感受野(Receptive Field)的机制而提出的一种前馈神经网络。卷积神经网络借助于卷积、池化等操作用于局部特征的提取[54]。 最初为计算机视觉发明，随后CNN模型被证明对NLP也同样有效，并在语义解析[55]，搜索查询检索[56]，句子建模[57]和其他传统的NLP任务都有不错的表现。

目前的卷积神经网络一般采用交替使用卷积层和最大值池化层，然后在顶端使用多层全连接的前馈神经网络。训练过程使用反向传播算法。卷积神经网络在结构上的特点有三个：局部连接，权重共享以及次采样。这些特点使得卷积神经网络在一定程度上具有平移不变性、缩放不变性和扭曲不变性。并且和前馈神经网络相比，卷积神经网络的参数更少，因而训练速度更快。在图像识别任务上，基于卷积神经网络模型的准确率也远远超出了一般的神经网络模型。

在全连接前馈神经网络中，假设第层有个神经元节点，第层有有个神经元节点，那么连接边有条，也就是权重矩阵有个参数。当和都很大时，权重矩阵参数将会非常多，训练的速度会非常得慢。如果采用卷积来代替全连接，即第层的每一个神经元都只和第层的一个局部窗口内的神经元相连接，从而构成一个局部连接网络。第层的第个神经元的输入可以定义为：

|  |  |
| --- | --- |
|  | (3-7) |
|  | (3-8) |

其中，为*m*维的滤波器，。这里，的下标从1开始。上述公式也可以写为:

|  |  |
| --- | --- |
|  | (3-9) |

表示卷积运算，从公式3-8可以看出，对于所有的神经元都是相同的。这也是卷积层的另外一个特性就是权值共享。这样，在卷积层里，只需要*m+*1个参数。另外第层的神经元个数不是任意选择的，而是满足。

卷积层虽然可以显著减少连接边的条数，但是对于每一个特征映射的神经元个数还是很多并没有明显减少。因此如果后面接一个分类器，分类器的输入维数依然很高，这样会很容易出现过拟合的情况。为了解决这个问题，在卷积神经网络当中一般会在卷积层之后再加上一个Pooling的操作，也就是子采样(Subsampling)，构成一个子采样层。子采样层可以大大降低特征的维数，从而避免过拟合情况的发生。对于卷积层得到的一个特征映射，可以将划分为很多区域，，这些区域可以重叠，也可以不重叠。一个子采样函数**down**(…)定义为：

|  |  |
| --- | --- |
|  | (3-10) |
|  | (3-11) |

其中，和分别是权重和偏置参数。子采样函数**down**(.)一般是取区域内所有神经元的最大值(Maximun Pooling)或者平均值(Average Pooling)。

|  |  |
| --- | --- |
|  | (3-12) |
|  | (3-13) |

子采样还有一个作用就是使得下一层的神经元对于上一层一些小的形态改变保持不变性，并且可以拥有更大的感受野。

图3-3是由Kim[58]提出的用于句子分类的CNN模型，它一共包含四层，分别是输入层、卷积层、池化层以及最后的输出层。输入层为一个的词向量矩阵，其中为词向量的维度，为句子所包含词语的个数。输出层为一个的向量，其中为类别数，得到的是句子在不同类别上的一个概率分布。

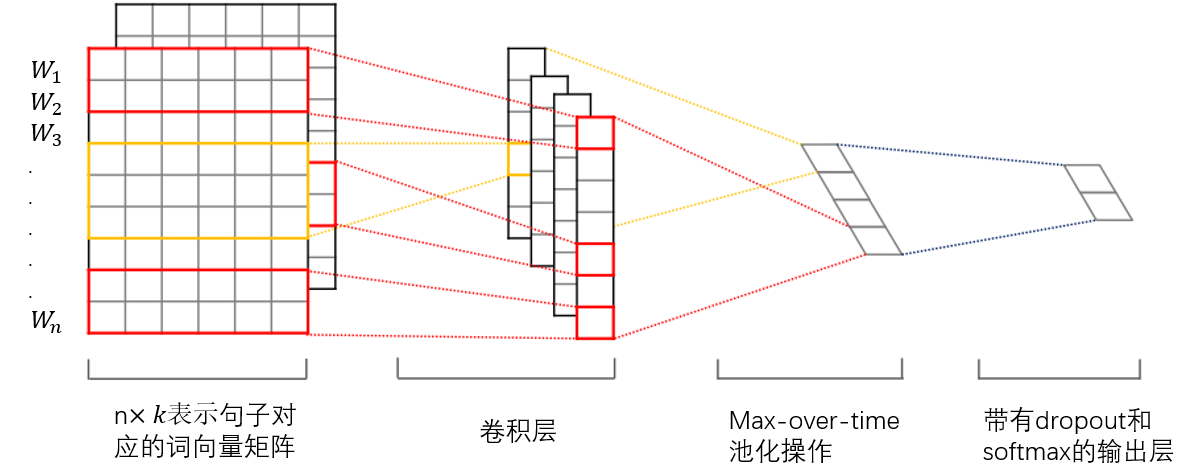


图3-3基于CNN的句子分类模型示意图

关系抽取是一个非常复杂任务，仅依靠词级别上特征还是不够的，一般还需要一些句子结构上的信息。因为有些关系是具有方向性的，比如投资关系：“2015年9月30日，SoFi通过官网声明软银当日以10亿美元领投SoFi”。这句话中包含了两个命名实体分别是“软银”和“SoFi”，它们之间具有投资关系：软银投资了SoFi。显而易见的是，软银是投资的主动发起方，而SoFi则是投资的接受方。这种单向的实体间关系使得关系抽取任务变得更加复杂，仅仅依靠上述模型无法区分出来关系的主动方与被动方。

为此，本文尝试在上述模型的基础上加入了词的位置信息，用于构建带有句子结构信息的分布式表示。具体实现过程如下：

* + 1. 位置嵌入

在关系抽取的任务中，一般距离实体词越近的词对于确定实体间关系的可能性越大。和Zeng的工作类似，本文使用由实体对指定的位置嵌入信息来帮助CNN网络记录每个词语与头尾实体的距离。它被定义为从当前单词到头实体和尾实体的相对距离的组合。举个例子，在句子“2015年9月30日，SoFi通过官网声明软银当日以10亿美元领投SoFi”当中，经过分词、去除停用词的预处理之后可以得到如下词序列：

{ 2015年**|** 9月**|** 30日**|** SoFi**|** 通过| 官网| 声明| 软银| 当日| 10亿| 美元| 领投| SoFi }

其中头实体为“软银”，尾实体为“SoFi”，“领投”这个词与头实体“软银”的距离为4，与尾实体“SoFi”的距离为1。假设表示位置嵌入，WE表示词向量，那么，其中分别表示句子当中某个词与头实体和尾实体的距离。将PE与WE结合在一起即可得到词的表示：，接着将其作为输入送入到卷积神经网络当中。

* + 1. 卷积和池化

假设表示句子当中第*i*个词语对应的*k*维词向量(本文*k*为200)和位置嵌入的组合，一个长度为*n*的句子则可以表示为：

|  |  |
| --- | --- |
|  | (3-14) |

其中，表示连接操作符，一般来说表示将词语,,…,连接之后得到的结果。之后用一个滤波器进行卷积操作，这个滤波器将与一个窗口大小为*h*词向量矩阵产生一个特征。特征可由包含词语的窗口通过如下公式产生：

|  |  |
| --- | --- |
|  | (3-15) |

这里代表偏置项，是一个非线性函数例如双曲正切函数。这个滤波器通过窗口滑动与句子当中所有词语卷积得到一个特征图谱，这里**。**然后对该特征图谱使用一个max-overtime的池化操作，即取作为此滤波器下得到的特征。很显然，这里做池化操作的目的就是处理不同长度的句子，这样无论句子长度为多少，卷积核宽度是多少，最终到得到定长的向量表示，同时max-pooling也是为了捕获最重要的特征信息。紧接着使用多个滤波器(具有不同的窗口大小)来获取多个不同的特征。这些特征作为倒数第二层，并被传递到全连接的softmax层，其输出则是类别上的概率分布。在具体实验中，窗口大小设置为3，4，5，每种窗口对应100种不同的滤波器将会对应产生100个特征图谱。

* + 1. Dropout和L2正则化

为了避免模型的过拟合，本文使用Dropout[59]和L2正则化技术[60]来加强模型的泛化性能。Dropout是一种相当激进的技术，和 L1、L2 正则化不同，Dropout并不依赖对代价函数的修改，而是改变了网络本身。Dropout操作会随机地使网络中的某些节点失效，即随机删除某些神经元，同时让输入层和输出层的神经元保持不变。

而L2正则或叫权重衰减(weight decay)是一种最常用的正则化技术，它的思想是在代价函数中加入一个额外的正则化项。公式3-15是正则化之后的交叉熵：

|  |  |
| --- | --- |
|  | (3-16) |

第一项是常规的交叉熵表达式,第二项是正则化项，也就是网络中所有权值的平方和。它由参数进行调整，其中被称为*正则化参数(regularization parameter)*，是训练集的大小。直观来说，正则化的作用是让网络偏好学习更小的权值，而在其它的方面保持不变。选择较大的权值只有一种情况，那就是它们能显著地改进代价函数的第一部分。换句话说，正则化可以视作一种能够折中考虑小权值和最小化原来代价函数的方法。具体实验中，本文Dropout率设置为0.5， 设置为3，mini-batch大小设置为50。

* + 1. 反向传播训练

反向传播算法最早于上世纪70年代被提出，但是直到1986年，由David Rumelhart, Geoffrey Hinton, 和Ronald Williams联合发表了一篇著名论文之后[61]，人们才完全认识到这个算法的重要性。文章介绍了几种神经网络，在这些网络的学习中，反向传播算法比之前提出的方法都要快，这使得以前用神经网络不可解的一些问题，现在可以通过神经网络来解决。反向传播算法是神经网络学习过程中的关键所在。

在本文中，训练样本中的每个句子都是相互独立的。对于给定一个样本*s*，输入到网络中，与训练参数运算将输出一个向量，它的第*i*维表示该句子属于第*i*类的一个概率评分。为了获得条件概率,我们在关系类别之上做一个softmax操作:

|  |  |
| --- | --- |
|  | (3-17) |

对于所有的训练样本*T*:(,)，我们可以得到参数的对数似然函数值：

|  |  |
| --- | --- |
|  | (3-18) |

为了计算网络的参数，我们使用随机梯度下降(SGD)技术来最大化似然函数。由于参数分布在网络的不同层当中，因此我们通过反向传播算法来迭代更新：

|  |  |
| --- | --- |
|  | (3-19) |

* 1. 对比实验

为了证明两种句子向量化方式的有效性以及比较两者优劣，本文做了一个文句分类的对比实验。具体实验结果参照3.4.3小节。

* + 1. 数据集及评价标准

本章实验数据来自SemEval-2010 Task 8[1]会议所给出标准数据集，该数据集一共包含了10717条数据，定义了9种实体关系，其中包括：1)Cause-Effect；2)Component-Whole；3)Entity-Destination；4)Entity-Origin；5)Product-Producer；6)Member-Collection；7)Message-Topic；8)Content-Container；9)Instrument-Agency。其中Other用于表示训练样本中的关系不属于九种主要关系类型中的任何一种。每条数据都包含一个标记有两个实体和的句子，最终任务是预测两个实体间的关系。

表3-1 语料统计

|  |  |
| --- | --- |
| **Relation** | **Freq** |
| Cause-Effect | 1331(12.4%) |
| Component-Whole | 1253(11.7%) |
| Entity-Destination | 1137(10.6%) |
| Entity-Origin | 974(9.1%) |
| Product-Producer | 948(8.8%) |
| Member-Collection | 923(8.6%) |
| Message-Topic | 895(8.4%) |
| Content-Container | 732(6.8%) |
| Instrument-Agency | 660(6.2%) |
| Other | 1864(17.4%) |
| Total | 10717(100%) |

其中Freq表示每一类关系语料的绝对数量以及占整个语料库的一个比重。

本文采用分类方法中常用的三种评价方法对两个算法进行评测：分别是准确率(Precision)、召回率(recall)和F值。它们的计算公式如下：

|  |  |
| --- | --- |
|  | (3-20) |
|  | (3-21) |
|  | (3-22) |

* + 1. 实验设置

1)本文选取了英文维基作为训练词向量的语料库，一共包含375万篇文章[[11]](#footnote-11)，经过简单的预处理操作后得到一个约11G大小的文本文件wiki.en.text。其中训练时所用的参数见表格3-2。

表 3-2 训练词向量参数

|  |  |
| --- | --- |
| 词向量维度 | 200 |
| 窗口大小 | 5 |
| 训练算法 | hierarchical softmax |
| 采样阈值 | 1e-3 |

训练结束后将得到一个所有词语的词向量矩阵，并以二进制形式存于文件wiki\_english\_word2vec(Google).model中，通过加载这个文件到内存中便可以得到指定词语的词向量以及计算词与词之间的语义距离。

2)卷积神经网络中滤波器窗口大小设置为3，4，5，每种窗口对应100种不同的滤波器将会对应产生100个特征图谱。Dropout率设置为0.5， 设置为3，mini-batch大小设置为50。实际训练中根据十折交叉法的结果来计算模型的准确率、召回率和F值。将总的数据集平均划分成十份，每次将其中的九份作为训练数据集，剩下的一份作为测试数据集。

表 3-2 CNN参数设置

|  |  |
| --- | --- |
| *h* | 3、4、5 |
| filter size | 100 |
| Drop rate | 0.5 |
| mini-batch | 50 |
| Learning rate | 0.01 |
| Hidden layer 1 | 200 |
| Hidden layer2 | 100 |

* + 1. 实验结果及分析

表3-3 基于词向量加权的朴素贝叶斯分类实验结果

|  |  |  |  |
| --- | --- | --- | --- |
| Precision | Recall | F-Measure | Relation |
| 0.619 | 0.661 | 0.639 | 1 |
| 0.701 | 0.724 | 0.712 | 2 |
| 0.749 | 0.727 | 0.738 | 3 |
| 0.723 | 0.680 | 0.701 | 4 |
| 0.716 | 0.707 | 0.711 | 5 |
| 0.667 | 0.652 | 0.659 | 6 |
| 0.727 | 0.626 | 0.672 | 7 |
| 0.653 | 0.768 | 0.706 | 8 |
| 0.705 | 0.781 | 0.741 | 9 |
| 0.695 | 0.702 | 0.698 | Avg |

表3-4 基于CNN的分类实验结果

|  |  |  |  |
| --- | --- | --- | --- |
| Precision | Recall | F-Measure | Relation |
| 0.719 | 0.661 | 0.689 | 1 |
| 0.801 | 0.824 | 0.812 | 2 |
| 0.849 | 0.827 | 0.838 | 3 |
| 0.823 | 0.780 | 0.801 | 4 |
| 0.916 | 0.907 | 0.911 | 5 |
| 0.767 | 0.852 | 0.807 | 6 |
| 0.827 | 0.826 | 0.826 | 7 |
| 0.753 | 0.768 | 0.760 | 8 |
| 0.805 | 0.781 | 0.793 | 9 |
| 0.806 | 0.802 | 0.798 | Avg |

其中Avg代表的是每一列的平均值。Relation一列当中的数字1、2、3…9分别对应上文提到的9种关系类型。根据两个表格中的F值可画出对比折线，如图3-4所示。

图3-4 两种方法的F值对比

实验结果图像表明两种向量化方式在文句分类方面的效果都很好，说明两种模型都能学习出文本的低维语义表示。通过对比表3-3和3-4的结果可以发现，基于CNN的句子分类方法无论是准确率还是召回率都要比词向量加权的方法略胜一筹。

* 1. 本章小结

在实现关系抽取的相关任务时，通常做法是将其转换为句子的分类任务，即事先定义好几种关系，然后构建关系的分类模型。而对于句子的分类任务来说，句子的语义向量表示是不可或缺的重要步骤。传统的词袋模型存在两个比较明显的缺陷：一是它丢失了词与词之间的顺序信息，二是它没有考虑词的语义信息。相比于词袋模型的高特征和语义缺失，我们希望能够生成紧凑和富有语义的句子向量。当前最热门、效果最好的语义向量化是RNN，但是结构比较复杂。

本文借助于词向量尝试使用了两种方法学习句子的紧凑的分布式表示，一种是词向量加权，将复杂的句子向量化(text-vector)问题转化为相对容易的词向量化(word-vector)问题，而词向量容易通过word2vec训练得到。另一种方法通过卷积神经网络来构建句子的分类模型，其中通过池化操作来处理不同长度的句子，这样无论句子长度为多少，卷积核宽度是多少，最终到得到定长的句子向量表示。

经过句子分类的实验论证，两种方法训练出来的语义向量表现良好,都可以学习出句子紧凑的分布式表示。其中基于CNN的句子分类算法在SemEval-2010 Task 8标准数据集上的表现要优于词向量加权的算法。

1. 面向互联网新闻文本的企业关系抽取
   1. 引言

互联网上每天都会更新产生大量的新闻信息，这些信息主要以网页文本的形式存在，其中所包含的企业以及人物之间的关系是一种有重要价值的信息。它在情报分析、网络舆情监控、社会网络分析等领域有着十分重要的应用。本章将在第三章的基础上，结合网页正文提取、命名实体识别、基于Bootstrapping的语料库构建等技术实现对中文新闻网页当中的企业关系进行抽取。

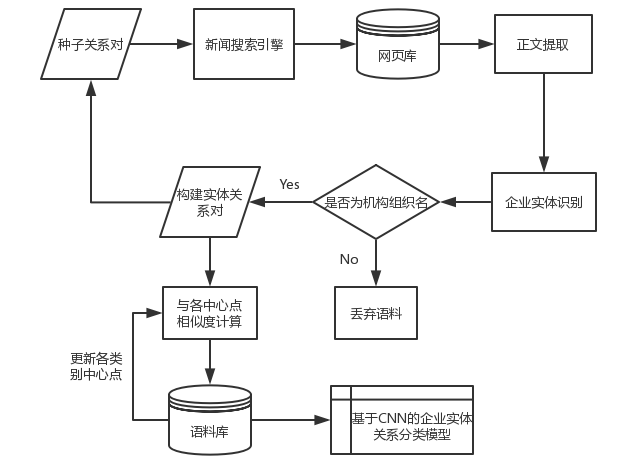


图4-1 面向互联网新闻文本的企业关系抽取系统框架

图4-1是整个系统的框架设计图，主要包括两大部分，分别是语料库构建模块以及关系分类模块。其中正文提取和企业实体识别在构建语料和关系抽取过程中都有涉及，属于公共部分。所以接下来在4.2和4.3小节将着重介绍相关技术的实现方法。4.4将介绍语料库的构建过程。

* 1. 网页正文提取

对于新闻网页而言，真正有用的信息一般只存在于新闻标题和正文当中，网页其它部分对于我们来说是一种噪音，例如广告栏、相关链接和一些脚本语言等。因此需要借助网页正文的提取技术将网页的正文抽取出来。针对这一问题，很多方法已经被提出来，常用方法有基于模板的方法、基于文字块密度的方法、基于DOM树节点统计的方法以及基于视觉的方法。

基于模板的网页正文提取算法依赖于网页HTML的内部结构上的一些特征。该方法假定同类型网页中有着相似的结构特征或者相似的Dom树(Document Object Model)结构。因而可以针对具有相似结构的网页来制定相同的模板来抽取网页的正文内容。举个例子，网易财经新闻频道当中的网页都有着相似的结构，正文都处在相同的div标签里。因此借助于HTML解析器，可以很容易地将所有该频道下网页的正文抽取出来。但这种方法的缺点也很明显，就是可移植性太差，需要针对不同结构的网页指定模板，相关工作费时费力。

基于文字块密度的方法是一种不需要建立DOM树的网页正文提取算法。该算法的主要思想是首先去除网页中所有的html 标签，然后根据去除 html 标签后的文字密度判断出正文区域。最后将所有的正文区域合并在一起，从而获得网页的正文内容。该方法只适用于结构相对简单的网页，对于一些相对复杂并且不规范的网页效果并不是很好。

基于建立DOM树的文档正文提取算法主要是先通过建立DOM树，然后根据DOM树中每个节点标签的文本数量和链接数、文本长度，并将链接文本的长度作为标准来确定点是否是正文节点。最后先序遍历DOM树，将所有的正文节点取出并整合在一起，完成正文提取功能。这种方法主要不足在于：1)需要建立网页的DOM树结构、时间复杂度比较高；2)不同网页之间差异性较大、很多网页编写得很不规范，比如标签丢失、不对齐等，加大了构建DOM树的难度。

上述两种方法对于网页结构复杂、正文区分布零散的网页提取效果不是很好，基于视觉的网页正文提取算法弥补了它们的不足。其中比较经典的基于视觉网页的提取算法是：VIPS(Visual Based Page Segment Algorithm)[62]，中文名为基于视觉的Web页面切割算。该算法的主要原理是从用户的视觉角度来分析网页的结构，即关心的是网页的视觉信息而不是网页的内部结构，它模拟人在阅读网页内容的过程并结合DOM树进行分析，可以弥补基于DOM树网页正文提取的一些不足之处，获得更加准确的抽取效果。图4-2为VIPS整个算法的流程图。该算法包含3个主要步骤：页面分块、块与块之间的权重计算、块的重构(组合)。这三个步骤组合在一起是一个完整的迭代过程。首先一个Web网页会被分割成几个大块并记录其层级结构。接着对于每个分割得到的大块对其进行处理，处理步骤即上述三个主要步骤。如此地递归处理下去直到获得足够多的小块区域并且达到某个阈值。

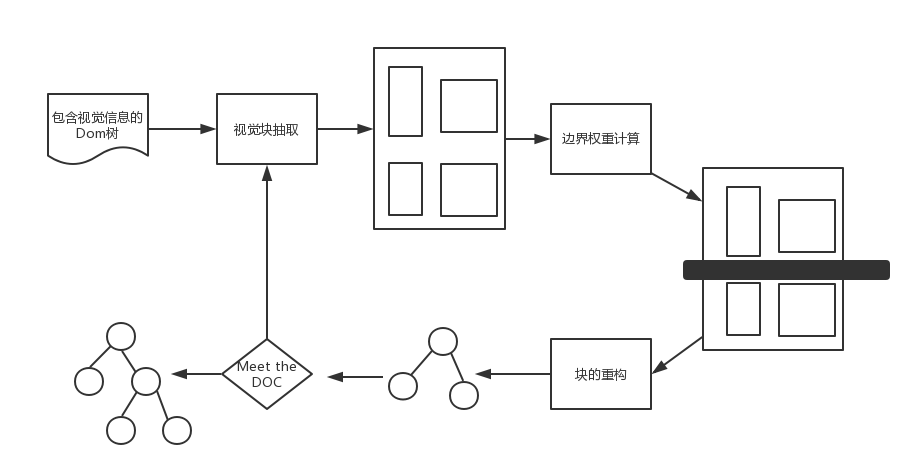


图4-2 VIPS算法流程图

基于视觉的网页正文提取算法充分利用网页的帧信息和视觉信息。与基于DOM树的网页文本提取算法相比，可以提高复杂结构和分散内容网页的提取准确性。然而，基于视觉的网页正文提取算法也有其固有的缺陷：首先，该算法需要多次迭代，最后需要语义块合并。与基于语义信息的网页正文提取算法相比，迭代更多，实现更复杂，增加时间复杂度。第二，提取网络视觉信息是耗时且费力的。由于网页的视觉信息采集和浏览器本身的层叠样式文件、脚本文件有关，访问视觉信息之前需要下载这些文件，接着浏览器内核就会调用这些文件，最后从浏览器的外部接口获取视觉页面信息此过程依赖于浏览器的内核代码并且非常耗时。

本文采用了哈工大的基于行块分布函数的通用网页正文抽取[63]所提出的算法，该算法将网页正文抽取问题转化为求页面的行块分布函数，并完全脱离HTML标签。通过线性时间建立行块分布函数图，由此图可以直接高效、准确的定位网页正文。同时采用统计与规则相结合的方法来解决系统的通用性问题。算法的依据有两点：1、正文区的密度；2、行块的长度。

依据1：一个网页的正文区域肯定是文字信息分布最密集的区域之一，这个区域可能最大但不尽然，比如评论信息较长，或者网页正文新闻较短，而又出现大篇紧密导航信息时。依据2：行块的长度信息可以有效解决上述存在的问题。

依据1和依据2相结合，就能很好的实现正文提取。算法流程如4.1所示：

|  |
| --- |
| 算法4.1：基于行块分布函数的通用网页正文抽取算法 |
| Input：原始网页HTML代码  output：网页正文内容 |
| 1. 首先将网页HTML标签去除干净，只留所有文字区域，同时保留标签去除后所有的空白位置信息，留下的正文成为Ctext； 2. 以Ctext中的行号为轴，取其周围行，合起来称为一个行块Cblock，行块是以Ctext中行号为轴的行块； 3. 对于一个Cblock，计算去掉其中的所有空白符(\n,\r,\t 等)后的字符总数做为该行块的长度； 4. 以Ctext每行为轴，共有个Cblock，画出以为横轴，以其各自的行块长度为纵轴的分布函数； 5. 根据上述行块分布函数图，找出骤升点和骤降点，这两个边界点所含的区域即为网页的正文。 |

行块分布函数可以在时间求得，在行块分布函数图上可以直观的看出正文所在的区域。如图4-3所示的网页行块函数分布图，正文区域行号为：140-181。

图4-3 行块函数分布图

* 1. 企业实体识别

命名实体识别是信息提取的子任务，其目的是找出文本中的命名实体的位置并将其分类为预定义的类别，例如人员、组织机构名、地点、时间表达、数量、货币价值、百分比等。

评判一个命名实体是否被正确识别主要从两个方面来考察：第一个是实体的边界是否正确；第二个是实体的类型是否标注正确。主要错误类型包括边界定位正确，类型可能判断错误；反之，文本边界错误，而其包含的主要实体词和词类标记可能正确。由于英文当中词与词之间天生存在间隔符，而且相关命名实体一般会有比较明显的标志，例如人名、地名等实体单词第一个字母都是大写等，所以实体识别相对于汉语来说要容易很多，因而任务的重点也就是确定实体的类型。和英文相比，汉语在识别命名实体之前需要先分词，而分词往往存在错误，因而使难度变得更大。

针对人名与机构名的自动识别问题，人们已经做过深入的研究，并提出了多种解决方法。其中主要技术方法分为：基于规则和词典的方法、基于统计机器学习的方法、以及规则与统计相结合的方法。

基于规则和词典的方法需要语言专家手动构建规则模板。特征包括统计信息、标点符号、关键字、位置字(如结束字)、中心词等。以字符串匹配为主要手段，这些系统大多依赖知识库和字典的建立。一般来说，当提取的规则可以更准确地反映语言现象时，基于规则的方法性能优于基于统计的方法。但是，这些规则往往依赖于特定的语言，领域和文本的风格，编写过程耗时耗费，难以覆盖所有的语言现象，尤其容易出错，系统可移植性差，对于不同的系统需要语言专家来重写规则。基于规则的方法的另一个缺点是成本太大，系统建设周期长，可移植性差，需要建立不同的领域知识库作为辅助，以提高系统识别能力。

基于统计的方法利用人工标注的语料进行训练，标注语料时不需要相关背景的语言学知识，并且可以在较短时间内完成。这类系统在移植到新的领域时不需要做太大的改动，只要构建新的语料库并进行一次训练即可。基于统计机器学习的方法主要包括：隐马尔可夫模型(Hidden Markov Model,HMM)[[12]](#footnote-12)、最大熵(Maximum Entropy,ME)[[13]](#footnote-13)、支持向量机(Support Vector Machine)、条件随机场(Conditional Random Fields,CRF)[[14]](#footnote-14)等。在这4种学习方法中，最大熵模型结构紧凑，具有较好的通用性，主要缺点是训练时间复杂性非常高，另外由于需要明确的归一化计算导致开销比较大。而条件随机场为命名实体识别提供了一个特征灵活、全局最优的标注框架，但同时存在收敛速度慢、训练时间长的问题。一般说来，最大熵和支持向量机在正确率上要比隐马尔可夫模型高一些，但是隐马尔可夫模型在训练和识别时的速度要快一些，主要是由于在利用Viterbi算法求解命名实体类别序列的时间效率较高。隐马尔可夫模型更适用于一些对实时性有要求以及像信息检索这样需要处理大量文本的应用，如短文本命名实体识别。

由于本文的研究重点是企业实体关系的抽取，因此本文直接采用了哈工大开源的语言技术平台(Language Technology Platform，LTP)[[15]](#footnote-15)来实现命名实体的识别。它的实现原理是将命名实体识别建模为基于词的序列标注问题。对于输入句子的词序列，模型给句子中的每个词标注一个标识命名实体边界和实体类别的标记。该平台支持人名、地名、组织机构名三类命名实体的识别。例如输入：

“2015年9月30日，SoFi通过官网声明软银当日以10亿美元领投SoFi”

返回结果为：

“2015年9月30日，[SoFi]Ns通过官网声明[软银]Ns当日以10亿美元领投[SoFi]Ns”

该方法主要是基于条件随机场实现的，条件随机场是一个无向图模型的框架，它能够被用来定义在给定一组需要标记的观察序列的条件下，一个标记序列的联合概率分布。假设分别表示需要标记的观察序列和它相应的标记序列的联合分布随机变量，那么就是一个以观察序列为条件的无向图模型。

我们定义为一个无向图，。即中的每个节点对应于一个随机变量所表示的标记序列的成分。如果每个随机变量对于G遵守马尔可夫属性(Markov property)，即前面所提到的条件独立属性，那么就是一个条件随机场，而且在给定和所有其他随机变量的条件下，随机变量的概率，即。

理论上，图G的结构可以是任意的，它描述标记序列中的条件独立性。然而，在建立模型时，最简单和最普遍的图结构是一个简单额一阶链结构。在该结构中的每个节点对应于的元素，的元素间并不存在图的结构，是因为我们只是将观察序列作为条件，并不对做任何的独立假设。

* 1. 基于Bootstrapping构建语料的方法

包含企业间关系的语料(这里主要是指句子)存在于互联网大大小小的新闻媒体网站当中，如果单靠人工逐个审查标注的方法，将会是一项费时费力的工程。本文借助于一些中文新闻搜索引擎(例如百度新闻搜索、360新闻搜索)，提出了一种基于Bootstrapping的方法用于搜集企业关系语料。算法描述如4-1所示。

|  |
| --- |
| 算法4-1：基于Bootstrapping构建语料的方法 |
| Input：初始关系种子集  output：包含企业关系的语料库 |
| 1. 构建关系类型体系； 2. 初始化种子关系集，形如<name(公司名)，relation(关系)>； 3. 根据种子集计算每一类关系的中心点； 4. 将name和relation作为关键字输入爬虫模块，得到一组包含公司名name与关系r的句子集合S; 5. 对于集合S当中的每个句子，逐个进行命名实体识别，找出其中出现的另一个公司名，得到实体对<name1,name2>; 6. 将上一步得到的实体对送入爬虫模块，得到候选语料集C 7. 针对候选语料C中的每一句子，计算其与每个类别中心点的距离，将距离最近的中心点的类标作为句子的类别； 8. 将(7)中的加入到对应类别的种子关系集中； 9. 判断语料库数量是否达到所设阈值，是进入(10)，否进入(3); 10. 结束 |

本文通过一个例子，来说明预料库具体的构建过程。

* + 1. 关系类型构建

企业与企业之间一般存在着投资(invest)、收购(acquisition)、合作(partner)、竞争(compete)的关系，企业与人一般存在的关系包括：董事长(chairman)、创始人(founder)。图4-1列举了一些知名企业之间的一些合作与竞争关系。从图中不难发现，有些公司之间存在一种以上关系，例如谷歌和苹果公司之间既存在竞争关系也存在合作关系。同时有些关系是单向的，有些关系是双向的。例如投资即为一种单向关系，A投资了B，反之并不成立。而有些关系是双向的，例如合作伙伴关系，A是B的合作伙伴，反过来B也是A的合作伙伴，在离散数学中，也叫这种关系具有对称性。

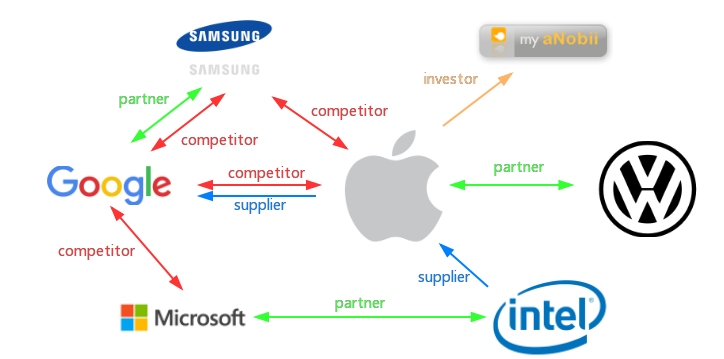


图4-1 企业之间的关系示意图

本文搜集整理100多家国内外企业公司名录，用于构建企业间关系的语料库。所选企业主要分为三大类：互联网IT类、汽车类以及传统制造业类。其中具有代表性的企业有：微软、谷歌、阿里巴巴、吉利、大众、格力电器等。针对这些企业，本文主要围绕投资(invest)、收购(acquisition)、合作(partner)、竞争(compete)、董事长(chairman)、其他 (NA)这6个关系从互联网上搜集相关语料。在标注语料的过程中，本文将区分头实体与尾实体，例如“软银投资了阿里巴巴”这句话中，“软银”是头实体而“阿里巴巴”是尾实体。而对应关系定义的方式，本文采取了对每一种关系定义一个相关关键词列表。这个关键词列表是由人工选择结合word2Vec扩展构建得到的。具体定义如表4-1所示。

表4-1 关系类型的定义及关键词列表

|  |  |
| --- | --- |
| **关系类型** | **关键词列表** |
| 合作(cooperate) | 合营 联合 中外合资 合资 协力 协同 协作 通力合作 合办 联手 联袂 携手 携手并肩 一并 一起 一同 分享 共享 共同 |
| 收购(acquisition) | 收购 并购 竞购 竞买 承购 购进 买进 买入 议购 函购 函售 卖 卖给 抛售 售卖 转售 贷款 营收 |
| 投资(invest) | 融资 投资 斥资 注资 投资额 竞得 投资者 入股 |
| 竞争(compete) | 竞争者 竞争 垄断 角逐 逐鹿 竞赛 比赛 |
| 董事长 | 董事 董事长 常务董事 董事局 执行主席 |

* + 1. 初始种子集构建

为了构建初始种子集，本文选取了各个行业中市值排名靠前的企业作为初始种子企业。国内互联网IT类的公司包括：腾讯、阿里巴巴、百度、蚂蚁金服、京东、网易、滴滴、携程、美团点评、今日头条。汽车行业类公司包括：丰田、大众、戴姆勒、宝马集团、本田汽车、通用汽车、福特汽车、日产汽车、现代汽车、上汽集团。制造行业类公司包括：格力电器、美的集团、飞利浦、海尔、索尼、松下电器、东芝、通用电器、夏普、三星集团。

接下来将上述企业名与表4-1中的关键词列表一一组合成种子关系对<entity1,relation-keyword> 。这样就可以得到一个关系对集合。我们将关系对中的企业名和关系词作为检索关键词与作为搜索引擎爬虫的输入，爬虫将自动爬取所有包含与的新闻网页。

对于持久化到本地的每个网页，将其送入网页正文提取器当中，得到该网页的正文，以句号和分号作为句子边界对正文进行划分得到句子集合。最后对集合进行筛选，只保留那些同时包含与的句子，筛选过后的句子集合为。

对于中的每个句子s，通过命名实体识别找出除以外的实体entity2。若entity2属于机构组织名，则将s进过布隆过滤器过滤判断后加入语料库中。且对于s来说，它的头实体为entity1，尾实体为entity2，关系为所对应的关系。同时将< entity1, entity2>作为关系对加入种子集合中。若entity2不属于机构组织名，则将该条预料丢弃。

* + 1. 句子的相似度计算及聚类

结合第二章提到的基于词向量加权的方法来获得句子的语义向量表示，我们将初始种子集中的每个句子进行向量表示，并借助K-Means聚类的思想对每类关系中的句子聚类求出中心点，用于之后计算待加入语料库句子类别相似度的评估。

文本向量的语义性在数学上主要体现在，语义越相似的文本之间的语义向量的距离也越近，计算方法主要包括欧式距离和夹角余弦：

|  |  |
| --- | --- |
| 欧式距离: | (4-1) |
| 夹角余弦： | (4-2) |

针对上小节构建的初始种子集，为了证明基于词向量加权能够很好的将同类句子聚集在一起，本文使用常见的高维数据可视化工具t-SNE[[16]](#footnote-16)对数据做了一个聚类展示。t-SNE(t-distributed stochastic neighbor embedding)是用于降维的一种机器学习算法，是由 Laurens van der Maaten 和 Geoffrey Hinton在08年提出来的，t-SNE 是一种非线性降维算法，非常适用于高维数据降维到2维或者3维，进行可视化。

t-SNE主要包括两个步骤：首先、t-SNE构建一个高维对象之间的概率分布，使得类似的对象有更高的概率被选择，而相似度较低的对象则有较低的概率被选择。其次，t-SNE在低维空间里在构建这些点的概率分布，使得这两个概率分布之间尽可能的相似，这里使用KL散度(Kullback–Leibler divergence)来度量两个分布之间的相似性。

t-SNE的降维原理就是让高维数据在映射到的低维空间中依旧保持和之前类似的分布。在原高维空间中两个点和的相似性定义为：

|  |  |
| --- | --- |
|  | (4-3) |

SNE是先将欧几里得距离转换为条件概率来表达点与点之间的相似度。具体来说，给定一个N个高维的数据(注意N不是维度)，t-SNE首先是计算概率，正比于和之间的相似度，即：

|  |  |
| --- | --- |
|  | (4-5) |

这里有一个参数是，对于不同的点取值不一样。此外设置，因为本文关注的是两两之间的相似度。其中是以点为中心的高斯分布的方差。而在映射后的低维空间，点和的相似性计算公式为：

|  |  |
| --- | --- |
|  | (4-6) |

这里使用了学生分布来衡量低维度下点之间的相似度。最后，我们使用KL散度来度量Q和P之间的相似度：

|  |  |
| --- | --- |
|  | (4-7) |

随后使用梯度下降来最小化KL离散度，梯度值如下所示：

|  |  |
| --- | --- |
|  | (4-8) |

通过优化算法可以使映射后的低维空间的相似度矩阵和原高维空间的相似度矩阵相近，此时得到的映射点就是降维后的数据。

算法的详细过程如下:

|  |
| --- |
| 算法4-1：用于降维可视化的t-SNE算法 |
| Input：高维数据表示  output：二维数据表示 |
| 1. Data：； 2. 计算cost function的参数：perplexity； 3. 优化参数：设置迭代次数T，学习率，动量; 4. 目标结果是低维数据表示; 5. 开始优化   计算在给定Perp下的条件概率(参见公式4-5)；  令；  用随机初始化  迭代，从到T，做如下操作：   1. 计算低维度下的(参见公式4-6) 2. 计算梯度(参见公式4-8) 3. 更新   *结束*   1. 结束 |

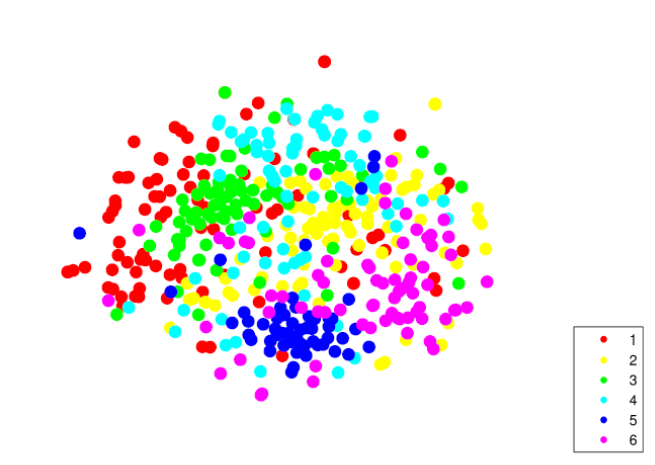


图4-2 基于t-SNE种子集聚类结果

* + 1. 语料库去重

由于网络新闻中存在大量转发、转载的情况，为了保证语料库的质量，本文在构建语料库的过程中会对其进行去重处理，防止同一个句子多次出现在语料库中。去重方法是基于布隆过滤器(Bloom Filter)的方法。

布隆过滤器是一种具有高效空间利用率的概率型数据结构，主要用于判断一个元素是否存在于集合中。相比较于哈希表，布隆过滤器只需要其1/8或1/4的空间复杂度就可以达到同样的效果。布隆过滤器可以插入元素，但不能删除已经存在元素。其中插入的元素越多，false positive rate(误报率)越大，但是不存在false negative (漏报)这种情况。

布隆过滤器的实现核心是一个超大的位数组和若干个哈希函数。不妨设位数组的长度为m，哈希函数的个数为k。

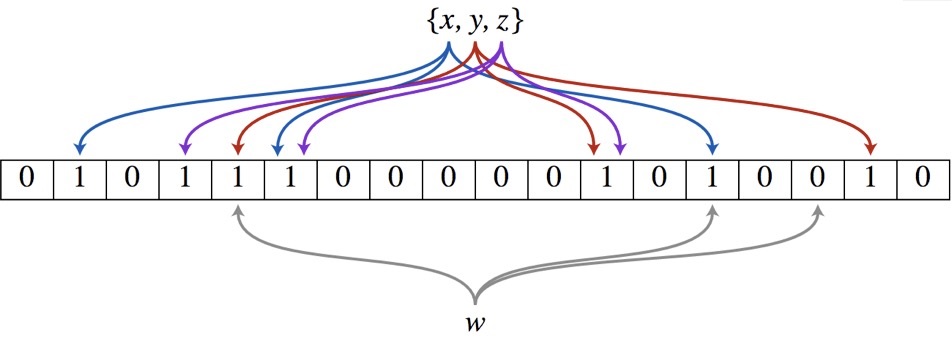


图4-3 布隆过滤器原理示意图

通过图4-3来说明布隆过滤器的具体操作流程：假设集合里有3个元素{x, y, z}，哈希函数的个数为3。第一步先将位数组进行初始化，将其每一位都设置为0。对于集合里面的每一个元素，将元素依次通过3个哈希函数进行映射，每次映射都会产生一个哈希值，这个值可以看成是数组的下标，即对应数组当中的某一位，然后将位数组对应的位置修改为1。当查询W元素是否存在集合中的时候，同样的方法将W通过哈希映射到位数组上的3个位置。如果这3个位置当中有一个不为1，则可以判断该元素一定不存在集合中。反之，如果3个点都为1，则该元素可能存在集合中。可以从图中可以看到：假设某个元素通过映射对应下标为4，5，6这3个点。虽然这3个点都为1，但是很明显这3个点是不同元素经过哈希得到的位置，因此这种情况说明元素虽然不在集合中，也可能对应的都是1，这是误判率存在的原因。

布隆过滤器添加元素的过程为：1)将要添加的元素计算出k个哈希函数值；2)得到对应于位数组上的k个位置；3)将这k个位置设为1。

布隆过滤器查询元素的过程为：1)将要添加的元素计算出k个哈希函数值；2)得到对应于位数组上的k个位置；3)如果k个位置有一个为0，则肯定不在集合中；4)如果k个位置全部为1，则可能在集合中。

* 1. 实验

在本章节中，本文主要在第三章基于CNN的关系分类模型的基础上实施了三个实验。第一个主要是调参，针对卷积神经网络中卷积窗口大小*h*和中间隐藏层的大小来考察它们对模型分类效果的影响；第二个实验主要是对比实验，主要和传统基于特征的关系抽取方法做了个对比。第三个实验是测试系统的实际表现效果，主要方法是从互联网中随机挑选一组包含企业关系的新闻URL，将其作为输入，将结果与人工抽取的结果做比较，得出准确率、召回率、F值。

* + 1. 数据集及评价标准

本章节实验数据来自于4.4小节所得到的中文新闻当中企业关系语料。共计1607条，其中各个关系类型语料分布如表4-1所示。

表4-1 语料统计

|  |  |
| --- | --- |
| **关系类型** | **频率** |
| 投资(invest) | 276(17.2%) |
| 收购(acquisition) | 291(18.1%) |
| 合作(cooperate) | 290(18.0%) |
| 竞争(compete) | 256(15.9%) |
| 董事长(chairman) | 293(18.2%) |
| NA | 201(12.5%) |
| Total | 1607(100%) |

整个系统的测试数据集为从互联网中随机选取的100篇包含企业关系的新闻，主要来源是网易新闻的科技频道(http://tech.163.com/)。其中所包含的企业实体以及企业实体间的关系事先已经标注好。

本章实验评价标准同第三章一样，分别是准确率(Precision)、召回率(recall)和F值。

* + 1. 实验设置

图4-4 不同窗口大小对应F值的变化

图4-5 隐藏层1大小对应F值的变化

图4-5 隐藏层2大小对应F值的变化

本小节主要通过实验来验证窗口大小和隐藏层大小对模型分类效果的影响。在图4-4和图4-5中，我们通过调整窗口大小*h*以及隐藏层大小*n*来计算不同参数下模型所对应的F值。最终结果是通过十折交叉验证得到的。从图中可以看出，当窗口大小为3时模型的F值最大，当在增加窗口大小时，模型的F值在变低。此外，由于我们使用的训练数据量较少，网络容易过拟合，特别是在使用较大的隐藏层时。从图4-5和图4-6可以看出，随着隐藏层的不断变大，对模型的影响变化很小。表4-2列出了本章实验的参数设置。

表 4-2 CNN参数设置

|  |  |
| --- | --- |
| Windows size | 3 |
| filter size | 100 |
| Drop rate | 0.5 |
| mini-batch | 50 |
| Learning rate | 0.01 |
| Hidden layer1 | 200 |
| Hidden layer1 | 100 |

* + 1. 实验结果及分析

为了和传统基于人工特征的方法做对比，本文做了一组对比实验，其中所选特征主要包括实体对、句子的依存句法树结构信息，实体对之间的词序列以及词性标注的信息。实验结果见表格4-3。

表 4-3 对比实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **分类器** | **特征集合** | **准确率** | **召回率** | **F值** |
| SVM | 词序列 | 0.652 | 0.613 | 0.632 |
| SVM | 词序列、词性标注 | 0.681 | 0.621 | 0.649 |
| SVM | 词序列、词性标注、依存句法树结构 | 0.694 | 0.701 | 0.697 |
| SVM | 词序列、词性标注、依存句法树结构、实体对位置信息 | 0.712 | 0.694 | 0.703 |
| **CNN** | 词向量、位置嵌入 | **0.786** | **0.752** | **0.768** |

由表4-3可见，基于特征的方法更多的是依赖于所选特征集合的质量，所提特征越多，模型的表现越好，但是提升空间有限。这种提升可以通过从训练到测试数据的语义泛化的需要来解释。传统特征的质量依赖于之前的经验和相关NLP工具，因此很难人工构建最优的特征子集。从最终结果来看，本文所提方法效果表现最好。

* 1. 本章小结

互联网上每天都会更新产生大量的新闻信息，这些信息主要以网页文本的形式存在，其中所包含的企业以及人物之间的关系是一种有重要价值的信息。它在情报分析、网络舆情监控、社会网络分析等领域有着十分重要的应用。本章将在第三章的基础上，结合网络爬虫、网页正文提取、命名实体识别等关键技术实现对中文新闻文本当中的企业、人物关系进行抽取。针对预料构建过程当中人工标注的效率低下的问题，提出了基于Bootstrapping技术构建关系语料库的方法，大大提高了获取语料库的效率。

1. 总结与展望
   1. 工作总结

本文主要工作围绕实体关系抽取任务的相关问题展开，主要包括两个部分，分别是句子的分布式表示和面向互联网新闻当中企业实体关系的抽取。需要注意的是，本文只考虑句子级别上的关系抽取。

第一部分针对传统词袋模型在表征句子时缺乏语义信息以及未考虑词的位置信息的缺陷，提出了基于词向量加权和基于卷积神经网络的方法用于构建紧凑且具有语义的句子分布式表示，作为构建关系分类模型的输入。

第二部分针对互联网新闻当中存在的企业实体关系，提出了基于Bootstrapping技术构建关系语料库的方法，克服了纯人工标注过程中费时费力的缺点。并在第一部分的工作基础上，结合网页正文提取、命名实体识别等关键技术，构建了面向互联网新闻文本的企业关系抽取系统的原型。

* 1. 未来展望

本文有关实体关系抽取的工作还存在着一些不足，比如目前只是考虑句子级的实体关系抽取，而没有考虑段落级和篇章级的实体关系抽取；另一方面本文主要面向的是互联网新闻领域的企业实体关系抽取，需要构建相关的语料库，无法直接迁移到别的领域。因此未来的工作主要专注于两个方面：

1. 文档级关系抽取：可以考虑引入等价关系，即共指消解处理结果，通过实体之间等价关系和非等价关系的融合和简单推理实现篇章级实体关系抽取，提高召回率，更好地对篇章进行理解。
2. 领域自适应的关系抽取：可以尝试构建一套领域自适应的关系抽取研究框架，即Open IE。系统可以自动发现关系类型、挖掘关系描述模式、抽取实体对。或者在已有领域标注语料库基础上，使用迁移学习(transfer learning)的方法推广到其他领域。

参考文献

1. Hendrickx I, Kim S, Kozareva Z. SemEval-2010 task 8: multi-way classification of semantic relations between pairs of nominals[C]. In *Proceedings of ACL*. Singapore ,2009.
2. Deng L, Yu D. Deep Learning: Methods and Applications[J]. *Foundations & Trends in Signal Processing*, 2014, 7(3):197-387.
3. Farabet C, Couprie C, Najman L. Learning Hierarchical Features for Scene Labeling[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2013, 35(8):1915-1929.
4. Dahl G E, Yu D, Deng L. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition[J]. *IEEE Transactions on Audio Speech & Language Processing*, 2012, 20(1):30-42.
5. Collobert R, Weston J, Bottou L, et al. Natural Language Processing (Almost) from Scratch[J]. *Journal of Machine Learning Research*, 2011, 12(1):2493-2537.
6. Bengio Y. Learning Deep Architectures for AI[J]. *Foundations & Trends® in Machine Learning*, 2009, 2(1):1-127.
7. Harris Z S. Distributional structure[M]. *Springer Netherlands:Papers in structural and transformational linguistics*, 1970: 775-794.
8. Hasegawa T, Sekine S, Grishman R. Discovering relations among named entities from large corpora[C]. In *Proceedings of ACL*,Barcelona,2004.
9. Chen J, Ji D, Tan C L, et al. Unsupervised feature selection for relation extraction[C].In *Proceedings of IJCNLP*. Jeju Island,Korea,2005.
10. Suchanek F M, Ifrim G, Weikum G. Combining linguistic and statistical analysis to extract relations from web documents[C]. In *Proceedings of SIGKDD*. Philadelphia, USA ,2006.
11. Qian L, Zhou G, Kong F, et al. Exploiting constituent dependencies for tree kernel-based semantic relation extraction[C]. *In Proceedings of COLING*. Manchester, UK ,2008.
12. Bunescu R C, Mooney R J. Subsequence Kernels for Relation Extraction[J]. *Advances in Neural Information Processing Systems*, 2005:171-178.
13. Bunescu R C, Mooney R J. A shortest path dependency kernel for relation extraction[C]. In *Proceedings of ACL.* Sydney, Australia ,2005.
14. Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data[C] In *Proceedings of ACL.* Singapore,2009.
15. Takamatsu S, Sato I, Nakagawa H. Reducing wrong labels in distant supervision for relation extraction[C] In *Proceedings of ACL*. Jeju Island, South Korea,2012.
16. Yao L, Riedel S, McCallum A. Collective cross-document relation extraction without labelled data[C]. In *Proceedings of ACL*.Uppsala,Sweden,2010.
17. Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text[J]. *Machine learning and knowledge discovery in databases*.2010: 148-163.
18. Hoffmann R, Zhang C, Ling X, et al. Knowledge-based weak supervision for information extraction of overlapping relations[C]. In *Proceedings of ACL.*Oregon,USA,2011.
19. Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance multi-label learning for relation extraction[C]. In *Proceedings of ACL*. Jeju Island, South Korea,2012.
20. Socher R, Huval B, Manning C D, et al. Semantic compositionality through recursive matrix-vector spaces[C]. In *Proceedings of ACL*. Jeju Island, South Korea,2012.
21. Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods[J]. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 1970:189--196.
22. Blum A, Mitchell T M. Combining Labeled and Unlabeled Sata with Co-Training[C]. *Eleventh Conference on Computational Learning Theory*, Wisconsin, Usa, 1998.
23. McDonald R. Extracting relations from unstructured text[J]. *Rapport technique*, *Department of Computer and Information Science*,University of Pennsylvania, 2005.
24. Abney S. Understanding the Yarowsky Algorithm[J]. *Computational Linguistics*, 2006, 30(3):365-395.
25. Brin S. Extracting Patterns and Relations from the World Wide Web[J]. *Lecture Notes in Computer Science*, 1998, 1590:172-183.
26. Agichtein E, Gravano L. Snowball : extracting relations from large plain-text collections[C]. In *Proceedings of  ACM Conference on Digital Libraries*, San Antonio, USA,2000.
27. Etzioni O, Cafarella M, Downey D, et al. Unsupervised named-entity extraction from the Web: An experimental study [J]. *Artificial Intelligence*, 2005, 165(1):91-134.
28. Banko M, Cafarella M J, Soderland S, et al. Open information extraction from the web[C]. In *Proceedings of* IJCAI. Hyderabad, India,2008.
29. Brin, Sergey. Extracting Patterns and Relations from the World Wide Web[M]. *The World Wide Web and Databases. Springer Berlin Heidelberg*, 1998:172-183.
30. Huang X, You H, Yu Y. A Review of Relation Extraction[J]. *New Technology of Library & Information Service*, 2013.
31. Kambhatla, Nanda. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations[C]. In *Proceedings of ACL*,Barcelona,2004.
32. Zhao S, Grishman R. Extracting relations with integrated information using kernel methods[C]. In *Proceedings of ACL*. Sydney, Australia,2005.
33. Miller S, Fox H, Ramshaw L, et al. A novel use of statistical parsing to extract information from text[C]. In *Proceedings of ACL*. Hong Kong, 2000.
34. Culotta A, Mccallum A, Betz J. Integrating probabilistic extraction models and data mining to discover relations and patterns in text[C]. In *Proceedings of ACL*. Sydney, Australia, 2006.
35. Lodhi H, Saunders C, Shawe-Taylor J, et al. Text classification using string kernels[J]. *Journal of Machine Learning Research*, 2002, 2(3):419-444.
36. Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network[C]. In *Proceedings of COLING*, Dublin, Ireland,2014.
37. Santos C N, Xiang B, Zhou B. Classifying relations by ranking with convolutional neural networks[J]. *Computer Science*, 2015, 30(3):365-395.
38. Miwa M, Bansal M. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures[C]. In *Proceedings of ACL*. Jeju Island, South Korea,2012.
39. Grishman R, Sundheim B. Message Understanding Conference-6: a brief history[C]. In *Proceedings of ACL*. California, USA,1996.
40. Humphreys K, Gaizauskas R, Azzam S, et al. In Proceedings of the Seventh Message Understanding Conference (MUC-7) [C]. In *Proceedings of MUC*. Fairfax, VA,1998.
41. Doddington G R, Mitchell A, Przybocki M A, et al. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation[C]. In *Proceedings of LREC*. Lisbon,Portugal,2004.
42. Alperin J L. Local representation theory[M]. *England:Cambridge University Press*, 1986.
43. Sahlgren M. The distributional hypothesis[J]. *Italian Journal of Linguistics*, 2008, 20(1): 33-54.
44. Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. *Computer Science*, 2013,30(3):365-395.
45. Le Q V, Mikolov T. Distributed Representations of Sentences and Documents[J]. *Computer Science*, 2014, 4:1188-1196.
46. Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks[J]. *Advances in Neural Information Processing Systems*, 2014, 4:3104-3112.
47. Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. *Advances in neural information processing systems*, 2013, 26:3111-3119.
48. Lebret R, Collobert R. Word Emdeddings through Hellinger PCA[J]. *Computer Science*, 2013, 4:3104-3112.
49. Levy O, Goldberg Y. Neural word embedding as implicit matrix factorization[J]. *Advances in Neural Information Processing Systems*, 2014, 3:2177-2185.
50. Li Y, Xu L, Tian F, et al. Word embedding revisited: a new representation learning and explicit matrix factorization perspective[C]. *In Proceedings of AAAI*, Austin,USA,2015.
51. Globerson A, Chechik G, Pereira F, et al. Euclidean Embedding of Co-occurrence Data[J]. *Journal of Machine Learning Research*, 2004, 8(4):2265-2295.
52. Levy O, Goldberg Y. Linguistic Regularities in Sparse and Explicit Word Representations[C]. In *Proceedings of COLING*. Dublin, Ireland,2014.
53. Nicholas Metropolis, S. Ulam. The Monte Carlo Method[J]. *Journal of the American Statistical Association*, 1949, 60(247):252.
54. Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. *Proceedings of the IEEE*, 1998, 86(11):2278-2324.
55. Yih W T, He X, Meek C. Semantic Parsing for Single-Relation Question Answering[C]. In *Proceedings of ACL*. Jeju Island, South Korea,2014.
56. Shen Y, He X, Gao J, et al. Learning semantic representations using convolutional neural networks for web search[C].Companion Publication of the, International Conference on World Wide Web Companion. 2014:373-374.
57. Kalchbrenner, E. Grefenstette, P. Blunsom. 2014. A Convolutional Neural Network for Modelling Sentences. In *Proceedings of ACL*.Maryland,USA,2014.
58. Kim Y. Convolutional Neural Networks for Sentence Classification[J]. *Eprint Arxiv*, 2014,8(4):2265-2295.
59. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. *Journal of Machine Learning Research*, 2014, 15(1):1929-1958.
60. Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. *Computer Science*, 2012, 3(4):212-223.
61. Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. *Parallel Distributed Processing Explorations in the Microstructure of Cognition*, 1986, 323(6088):533-536.
62. Cai D, Yu S, Wen J R, et al. VIPS: a Vision-based Page Segmentation Algorithm[J]. *Microsoft Research*, 2003.
63. 陈鑫. 基于行块分布函数的通用网页正文抽取[J].http://code. google. com/p/cx-extractor, 2010.

致谢

附录

研究生期间专利

1. 张雷,刘焕锐,资帅,王强,吴和生,谢俊元,一种中药方剂核心药物的发现方法:201510183745.0

研究生期间参与项目

1. 江苏银行贷后风险网络预警系统
2. 重庆市交委企业互联网信用数据采集与服务项目

附件二

《学位论文出版授权书》

本人完全同意《中国优秀博硕士学位论文全文数据库出版章程》(以下简称“章程” )，愿意将本人的学位论文提交“中国学术期刊(光盘版)电子杂志社”在《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》中全文发表。《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》可以以电子、网络及其他数字媒体形式公开出版，并同意编入《中国知识资源总库》，在《中国博硕士学位论文评价数据库》中使用和在互联网上传播，同意按“章程”规定享受相关权益。

作者签名：

2016 年 5 月 27 日

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 论文题名 | 基于语义重构的文本摘要算法 | | | | |
| 研究生学号 | Mg1333075 | 所在院系 | 计算机科学与技术 | 学位年度 |  |
| 论文级别 | □硕士 □硕士专业学位  □博士 □博士专业学位  (请在方框内画钩) | | | | |
| 作者Email | 870727617@qq.com | | | | |
| 导师姓名 | 王崇骏教授、吴骏讲师 | | | | |

论文涉密情况：

□ 不保密

□ 保密，保密期( 年 月 日 至 年 月 日)

注：请将该授权书填写后装订在学位论文最后一页(南大封面)。

1. [https:.en.wikipedia.org/wiki/Word\_embedding](https://en.wikipedia.org/wiki/Word_embedding) [↑](#footnote-ref-1)
2. [https:.en.wikipedia.org/wiki/DNN](https://en.wikipedia.org/wiki/DNN) [↑](#footnote-ref-2)
3. [https:.en.wikipedia.org/wiki/Freebase](https://en.wikipedia.org/wiki/Freebase) [↑](#footnote-ref-3)
4. <https://en.wikipedia.org/wiki/Co-training> [↑](#footnote-ref-4)
5. <https://en.wikipedia.org/wiki/String_kernel> [↑](#footnote-ref-5)
6. [https:.en.wikipedia.org/wiki/One-hot](https://en.wikipedia.org/wiki/One-hot) [↑](#footnote-ref-6)
7. [https:.en.wikipedia.org/wiki/Bag-of-words\_model](https://en.wikipedia.org/wiki/Bag-of-words_model) [↑](#footnote-ref-7)
8. <https://en.wikipedia.org/wiki/Word_embedding> [↑](#footnote-ref-8)
9. <https://code.google.com/p/word2vec/> [↑](#footnote-ref-9)
10. <https://en.wikipedia.org/wiki/Convolutional_neural_network> [↑](#footnote-ref-10)
11. <https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2> [↑](#footnote-ref-11)
12. <https://en.wikipedia.org/wiki/Hidden_Markov_model> [↑](#footnote-ref-12)
13. <https://en.wikipedia.org/wiki/Principle_of_maximum_entropy> [↑](#footnote-ref-13)
14. <https://en.wikipedia.org/wiki/Conditional_random_field> [↑](#footnote-ref-14)
15. <http://www.ltp-cloud.com/> [↑](#footnote-ref-15)
16. <https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding> [↑](#footnote-ref-16)