# 说 明 书 摘 要

本发明提供一种基于卷积神经网络的企业实体关系抽取的方法，包括如下步骤：1)构建关系语料库阶段：a)人工构建初始种子关系对；b)借助于互联网新闻搜索引擎并利用Bootstrapping技术迭代生成关系语料；2)关系分类模型训练阶段：a) 结合词向量与位置嵌入构建句子的向量矩阵表示作为网络的输入；b) 搭建卷积神经网络并利用反向传播算法训练网络得到关系分类模型；3) 新闻网页中企业实体关系抽取阶段：a)新闻网页预处理；b)对预处理过后的新闻网页中的企业实体关系进行抽取；4)结束。本发明使用了卷积神经网络来实现实体关系的抽取，避免了过多依赖于人工选取特征的方法。而基于人工特征的方法一般会存在两点不足：一是所提特征需要借助一些自然语言处理工具，这些工具难免会存在误差；二是难以选取合适的特征子集。通过本发明则可以避免上述问题，并可以准确高效地从新闻网页中抽取出企业实体关系。

# 摘 要 附 图



# 权 利 要 求 书

1、一种基于卷积神经网络的企业实体关系抽取的方法，其特征在于，包括构建关系语料库阶段、关系分类模型训练阶段、新闻网页中企业实体关系抽取阶段，具体包括如下步骤：

1)构建关系语料库阶段：

1a)人工构建初始种子关系对；

1b)借助于互联网新闻搜索引擎并利用Bootstrapping技术迭代生成关系语料；

2)关系分类模型训练阶段：

2a)结合词向量与位置嵌入构建句子的向量矩阵表示作为网络的输入；

2b)搭建卷积神经网络并利用反向传播算法训练网络得到关系分类模型；

3)新闻网页中企业实体关系抽取阶段：

3a)结合网页正文提取和命名实体识别技术对新闻网页进行预处理；

3b)对预处理过后的新闻网页进行企业实体关系进行抽取；

4)结束。

2、根据权利要求1所述的一种基于卷积神经网络的企业实体关系抽取的方法，其特征在于，步骤1a)中所述的人工构建初始种子关系对具体过程如下：

1)定义关系类型及整理对应的关键词列表；

2)选取若干上市企业名单作为初始种子企业；

3)将1，2进行两两组合得到初始种子关系对,例如<阿里巴巴，投资>；

4)结束。

3、根据权利要求1所述的一种基于卷积神经网络的企业实体关系抽取的方法，其特征在于，步骤1b)中所述的借助于互联网新闻搜索引擎并利用Bootstrapping技术迭代生成关系语料具体过程如下：

1)将1a)得到的种子关系对集合中企业名和关系关键词作为检索关键词与送入搜索引擎爬虫；

2)持久化同时包含与的所有网页；

3)对上一步得到的网页进行正文提取、分句等预处理操作；

4)筛选并保留同时包含与的句子作为初始关系语料S；

5)根据集合S计算每一类关系的中心点；

6)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<>，并将加入到种子企业中；

7)将上一步得到的实体对送入搜索引擎爬虫，得到候选语料集C；

8)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,并加入语料库S中，更新；

9)判断语料库数量是否达到阈值，若是则进入10)，否则进入1)；

10)结束。

4、根据权利要求1所述的一种基于卷积神经网络的企业实体关系抽取的方法，其特征在于，步骤2a)的具体过程如下：

1)利用word2vec训练得到每个词的词向量表示；

2)对于句子当中的每个词，计算其与两个实体的相对距离;

3)结合1、2得到每个词的语义与位置信息的组合表示；

4)则一个长度为n的句子则可以表示为：，表示连接操作符；

5)结束。

5、根据权利要求1所述的一种基于卷积神经网络的企业实体关系抽取的方法，其特征在于，步骤2b)的具体过程如下：

1)将2a)所得到的句子向量矩阵作为输入送入卷积神经网络中；

2)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图谱；其中为词向量维度+2；

3)对上步特征图谱使用一个max-overtime的池化操作，即取作为此滤波器下得到的特征；

4)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图谱并做池化操作得到相关特征；

5)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；

6)为了获得条件概率,我们在关系类别之上做一个softmax操作：；

7)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：；

8)通过反向传播算法来迭代更新：，并最终得到关系分类模型；

9)结束。

6、根据权利要求1所述的一种基于卷积神经网络的企业实体关系抽取的方法，其特征在于，步骤3a)中使用了开源工具WebCollector[[1]](#footnote-1)、HanLP[[2]](#footnote-2)分别实现网页的正文提取和命名实体识别任务。

7、根据权利要求1所述的一种基于卷积神经网络的企业实体关系抽取的方法，其特征在于，步骤3b)对预处理过后的新闻网页进行企业实体关系进行抽取具体过程如下：

1)将网页正文分句，主要以句号和分号作为分隔符；

2)筛选出包含两个或两个以上企业实体的句子集合S；

3)将S中的句子逐个送入关系分类模型中，得到该句子所包含的实体关系；

4)结束。

# 说 明 书

一种基于卷积神经网络的企业实体关系抽取的方法

**技术领域**

本发明涉及深度学习与自然语言处理，具体而言涉及一种基于卷及神经网络的实体关系抽取的方法。

**背景技术**

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。

实体关系抽取研究正是在这种背景下产生的。实体关系抽取作为自然语言处理的一项经典任务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。实体关系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，针对互联网新闻中存在的企业实体关系的抽取是一项比较典型的实体关系抽取任务，它对金融舆情、企业投资等方面具有重要作用。

目前，大多数实体关系抽取的方法是基于监督学习的方法，通常需要大量人工标注的训练数据，选取合适的特征后从训练数据中学习关系对应的抽取模式。该方法的缺点也很明显，一是需要人工标注训练数据集，十分的耗时耗力；二是依赖一些自然语言处理工具来提取特征，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。

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。借助于词向量和深层的神经网络例如卷积神经网络，可以学习得到句子的分布式向量表示。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。

**发明内容**

本发明目的在于提供一种基于卷积神经网络的企业实体关系抽取的方法，首先利用Bootstrapping技术构建企业关系语料库，接着基于卷积神经网络来构建句子的分布式表示，最后构建关系分类模型并对新闻网页中存在的企业关系进行抽取。

具体方法包括下列步骤：

1)构建关系语料库阶段：

1a)人工构建初始种子关系对；

1b)借助于互联网新闻搜索引擎并利用Bootstrapping技术迭代生成关系语料；

2)关系分类模型训练阶段：

2a)结合词向量与位置嵌入构建句子的向量矩阵表示作为网络的输入；

2b)搭建卷积神经网络并利用反向传播算法训练网络得到关系分类模型；

3)新闻网页中企业实体关系抽取阶段：

3a)结合网页正文提取和命名实体识别技术对新闻网页进行预处理；

3b)对预处理过后的新闻网页进行企业实体关系进行抽取；

4)结束。

本发明有益效果：本发明使用了卷积神经网络来实现实体关系的抽取，避免了过多依赖于人工选取特征的方法。而基于人工特征的方法一般会存在两点不足：一是所提特征需要借助一些自然语言处理工具，这些工具难免会存在误差；二是难以选取合适的特征子集。通过本发明则可以避免上述问题，并可以准确高效地从新闻网页中抽取出企业实体关系。

**附图说明**

图1为本发明基于卷积神经网络的企业实体关系抽取的方法的流程图；

图2为基于Bootstrapping技术的构建关系语料库的流程图；

图3为关系类型及对应关键词列表；

图4为构建基于卷积神经网络的关系分类模型的示意图；

图5为对新闻网页进行企业实体关系进行抽取的流程图。

**具体实施方式**

为了更了解本发明的技术内容，特举具体实施例并配合所附图式说明如下。

图1为本发明基于卷积神经网络的企业实体关系抽取的方法的流程图，如图所示，该方法主要包括三个阶段，分别是：基于Bootstrapping技术构建语料库阶段；基于卷积神经网络构建关系分类模型阶段；针对新闻网页进行企业实体关系抽取阶段。

本发明基于Bootstrapping技术构建语料库具体过程如图2所示，实施步骤如下：

步骤0为本发明的基于Bootstrapping技术构建语料库的起始状态；

步骤1为定义关系类型及整理对应的关键词列表，如图3所示；

步骤2选取若干上市企业名单作为初始种子企业；

步骤3将关系类型对应的关键词与种子企业名单进行两两组合得到初始种子关系集；

步骤4将种子关系集中企业名和关系关键词作为检索关键词与送入搜索引擎爬虫；

步骤5持久化同时包含与的所有网页；

步骤6对上一步得到的网页进行正文提取、分句等预处理操作；

步骤7为筛选并保留同时包含与的句子作为初始关系语料S；

步骤8是根据集合S计算每一类关系的中心点；

步骤9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<>，并将加入到种子企业中；

步骤10是将上一步得到的实体对送入搜索引擎爬虫，得到候选语料集C；

步骤11是针对候选语料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,并加入语料库S中，更新；

步骤12是判断语料库数量是否达到阈值，若是则进入步骤13，否则进入步骤3；

步骤13是构建关系语料库的结束状态。

图4为构建基于卷积神经网络的关系分类模型的示意图，其具体步骤为：

步骤14为起始步骤；

步骤15是利用word2vec训练得到每个词的词向量表示；

步骤16对于句子当中的每个词，计算其与两个实体的相对距离;

步骤17结合1、2得到每个词的语义与位置信息的组合表示；

步骤18对于一个长度为n的句子则可以表示为：，表示连接操作符；

步骤19将上一步得到的句子向量矩阵作为输入送入卷积神经网络中；

步骤20是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图谱；其中为词向量维度+2；

步骤21是对上步特征图谱使用一个max-overtime的池化操作，即取作为此滤波器下得到的特征；

步骤22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图谱并做池化操作得到相关特征；

步骤23是对于给定一个样本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；

步骤24为了获得条件概率,在关系类别之上做一个softmax操作：；

步骤25对于所有的训练样本T:(,)，可以得到参数的对数似然函数值：；

步骤26是通过反向传播算法来迭代更新：，并最终得到关系分类模型；

步骤27为构建关系分类模型的结束状态；

图5为对新闻网页进行企业实体关系进行抽取的流程图，其具体步骤为：

步骤28为起始状态；

步骤29是持久化网页到本地；

步骤30是借助WebCollector开源工具对网页进行正文提取；

步骤31是将网页正文分句，主要以句号和分号作为分隔符；

步骤32是借助HanLP开源工具对句子进行命名实体标注；

步骤33是筛选出那些包含2个或2个以上企业实体的句子；

步骤33将S中的句子逐个送入关系分类模型中，得到该句子所包含的实体关系类型；

步骤34为本发明的结束状态。

综上所述，本发明就是通过卷积神经网络来构建包含企业实体关系句子的分类模型，借助词向量和位置嵌入来构建带有语义和位置信息的句子向量表示，作为卷积神经网络的输入。这样做不但避免过多依靠人工特征的方法，还可以最大限度地减少一些自然语言处理工具所带来的误差，从而提高分类模型的准确性，具有良好的实用性。

虽然本发明已以较佳实施例揭露如上，然其并非用以限定本发明。本发明所属技术领域中具有通常知识者，在不脱离本发明的精神和范围内，当可作各种的更动与润饰。因此，本发明的保护范围当视权利要求书所界定者为准。

# 说 明 书 附 图



图1



图2

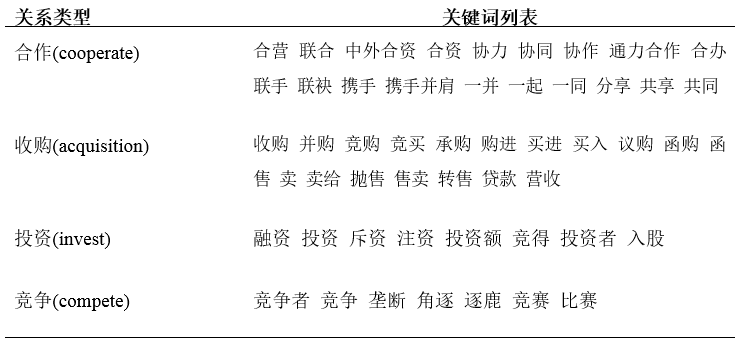


图 3

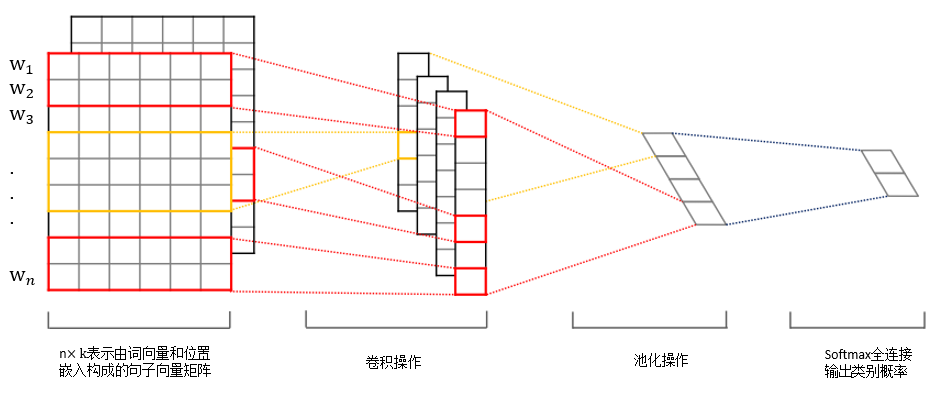


图4



图 5

1. <https://github.com/CrawlScript/WebCollector> [↑](#footnote-ref-1)
2. <https://github.com/hankcs/HanLP> [↑](#footnote-ref-2)