基于特征学习的文本大数据内容理解及其发展趋势

袁书寒，向 阳，鄂世嘉
同济大学计算机科学与技术系 上海 201804

摘要
大数据中蕴含着重要的价值信息，文本大数据作为大数据的重要组成部分，是人类知识的主要载体。特征作为数据内在规律的反映，将文本大数据映射到反映数据本质的特征空间是文本大数据语义理解的重要手段。介绍了文本大数据的特征表示、特征学习，进而梳理了特征学习在文本大数据内容理解中的进展，最后阐述了基于特征学习的文本大数据内容理解未来的发展趋势。

关键词
文本大数据；特征学习；内容理解

doi: 10.11959/j.issn.2096-0271.2015030

Text Big Data Content Understanding and Development Trend Based on Feature Learning

Yuan Shuhan, Xiang Yang, E Shijia
Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

Abstract
Big data contains important value information. Text big data as an important part of big data is the main carrier of human knowledge. Feature represents the inherent law of the data. Mapping the text big data to its feature space which reflects the nature of data is an important method to understand the semantic meaning of the text. Text big data feature representations and feature learning were reviewed. Then the progress of feature learning used in text content understanding was presented. Finally, the future development trends of big text data content understanding were discussed.

Key words
text big data, feature learning, content understanding

1 引言

近年来，随着互联网、云计算、社交网络的发展，网络空间中的信息总量在飞速膨胀，网络大数据时代已经到来。如何充分挖掘大数据中蕴含的价值成为全社会共同关注的话题。

在20世纪90年代，数据仓库之父比尔·恩门（Bill Inmon）提出数据仓库的概念，激活了沉睡在数据库中多年的历史数据，使之用于数据分析与决策支持，以挖掘出隐藏在数据背后的有价值信息。而在大数据时代，互联网每分钟都在产生大量的数据，YouTube每分钟内上传的视频长达72 h，Facebook上每分钟共分享了多达246万条信息，Instagram每分钟可产生21万张新照片[1]；在数据快速增长、数据类型多样、数据结构复杂的背景下，传统的基于静态、浅层的数据分析方法，已经无法适应当前越来越多的对数据语义深层理解和计算应用的需求。因此，大数据的分析、挖掘成为学术界、工业界共同的研究热点。

文本大数据是网络大数据的重要组成部分，人们日常工作和生活中接触最多的电子文档也是以文本的形式存在。从海量文本数据中挖掘有价值的信息、知识，一直都是学术界研究的热点问题，但是文本大数据的复杂性和规模性，导致传统的全量数据模式下对文本进行分析变得异常困难。挖掘海量文本数据的特征是降低计算时空复杂性、实现文本语义理解的重要手段。

本文主要介绍近年来伴随特征学习技术的发展，对海量文本数据特征发现，进而实现语义理解方面所取得的新进展。

2 文本大数据特征

人类是通过识别出物体的特征来认识不同的物体的，因此，特征作为数据本质的反映是理解数据的重要手段。将文本大数据映射到其特征空间，首先需要确定文本大数据的特征表示方式，正如不同的人认识同一物体时，会以不同的方式抽象物体的特征，特征表示方式也不尽相同，但是一个良好的特征表示方式是保证特征可理解、可计算的基础；在确定了特征表示方式的基础上，从文本大数据中学习能够精确表达文本语义的特征是实现内容理解的关键。

2.1 特征表示

由于文本大数据的多源异构性，实现海量文本的内容理解首先需要将非结构化的文本数据转化为计算机可操作的结构化模型，文本特征表示将文本信息映射到计算机可理解的特征空间，从而为计算机理解文本语义提供基础。在文本数据分析领域，传统的算法依赖人工定义反映输入数据性质的特征作为模型的输入，而为了尽可能地反映自然语言规律，提高算法的准确性，人工定义特征往往数量十分庞大，通常这一步骤称作特征工程。为了生成大量的特征，特征工程首先定义一系列的特征模板（feature template），利用特征模板进一步产生语言的特征。例如，在语言模型的研究中，定义三元特征模板（trigram feature template），对于训练语料库中出现的任意三元组（u, v, w），若在出现词语u、v的情况下，出现词语w，则该特征为1；类似地，还可以定义二元特征模板、一元特征模板或词语前缀模板等。

1 https://www.domo.com/learn/data-never-sleeps-2

从特征模板的定义可以看出，最终生成的特征可以高达数十万甚至数百万级别，这也导致人工定义的特征十分稀疏，只有极少部分的特征为非0值，而当测试语料中出现训练数据中没有的特征时，将训练数据产生的特征应用于测试数据，效果并不理想；且人工定义特征在面对特定任务时，通常存在过度细化的问题，而面对海量数据时，又存在表示不足的问题。

近年来，表示学习（representation learning）或非监督的特征学习（unsupervised feature learning）由于其可以自动地发现数据特征，从而有效地避免繁琐的人工参与，成为重要的研究方向。深度学习作为特征学习的主要手段，不仅可以利用海量训练数据实现分类、回归等传统机器学习的目标，还可以在模型的训练过程中产生层次化的抽象特征，该特征表示是提高训练准确性的重要基础。图1[1]对比了传统机器学习和深度学习在实现步骤上的不同。

一般而言，特征学习的目的在于学习一种数据的转换方式，用于从数据中抽取有效的特征信息，最终使得数据的分类、预测更加准确，而有价值的特征信息应该满足表达性、抽象性、排他性等要求[2]。

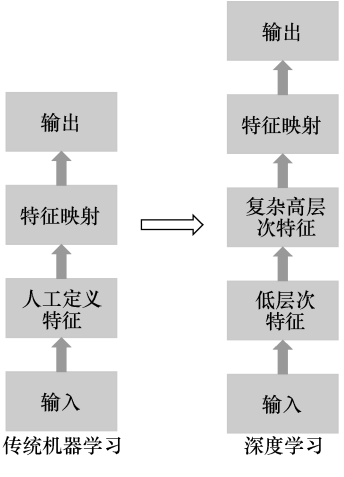


图1 深度学习与传统机器学习步骤对比

（1）表达性

表达性是指合理大小的特征应该能够有效表示足够大的输入数据。传统的文本数据理解以one-hot的形式表示，n维的空间只能表示n个特征。分布表示（distributed representations）是一种基于神经网络的表示方式，其思想来自于认知表示，它认为脑中的一个物体可以用许多描述该物体的神经元来有效表示，这些神经元可以独立地激活或不激活，例如，一个n维的二值神经元集合，可以描述 2^n个不同的数据，即每一个数据都由所有的神经元共同表示，而每个神经元都参与到各个不同数据的表示[3]中去。因此，分布表示可以看作由n维连续实值向量构成的特征空间，向量的每一维共同构成数据的特征表示向量，特征表示维度不会随着数据数量的增加而增加。

（2）抽象性

文本特征是对文本数据本身的抽象表示，因此文本的特征对文本数据的抖动应具有相应的顽健性，同时也不应该因任务的不同而变化。通常而言，对特征的抽象也具有层次性，低层次的抽象特征来源于输入数据，高层次的抽象特征来源于对低层次抽象特征的进一步学习，抽象的层次越高对数据抖动的不变性就越强，例如，相似的词汇、同义语句应该有相似的特征。因此，特征的抽象性反映了特征的不变性和层次性。

（3）排他性

文本特征的排他性是指特征应该刻画数据不同方面的性质，对于互不相同的性质，其特征也应该互相排斥。例如，文本是由文本的结构、文本中词语的选择、文本词语出现的顺序等多种互相关系的因素共同组成，而有效的特征表示应该能够尽可能多地分离出互相关联的因素，使得不同的抽象特征反映不同的文本内在因素。

文本大数据特征的表达性、抽象性和排他性定义了特征表示的不同层次，逐层递进。文本大数据的表达性保证了文本特征必须适合刻画非结构化数据，并且特征表示本身能够以固定的结构描述文本；在此基础上，文本特征应该是对文本内容的归纳和抽象，文本大数据是无穷尽的，但是特征应该是有限的；最后，特征的排他性要求特征能够使一个对象区别于其他对象，即如果一个文本具有某个特征，那么这个特征就能使这个文本区别于不具有这个特征的文本，从而为文本内容的精确理解提供基础。

2.2 特征学习

特征表示规约了特征的抽象形式，特征学习则指在选择特征表示的规范下，学习数据的特征。目前，对特征的学习主要有两类方法：一类是通过监督学习的方法，利用训练数据构建适合描述数据特征的模型；另一类是非监督学习的方法，该类方法主要通过降维将数据约简至特征空间，以发现数据的内在规律。近年来，由于深度学习可以自动发现结构化深层次特征，从而逐渐成为特征学习的主要方法。深度学习本质上是一个深度、多层的神经网络模型，由于它在图像处理、语音识别、自然语言处理等应用上的重大突破而成为研究热点。

2006年，Hinton等人[4]利用受限玻尔兹曼机（restricted Bolzman machine）非监督地预训练（pre-training）深层神经网络中每一层模型的参数，进而利用反向传播算法有监督地更新整体模型的参数，极大地提高了模型在图像识别上的准确率。其中，每一层受限玻尔兹曼机预训练得到的模型都可以看作对图像不同层次上的抽象特征。因此，早期的深度学习算法可分为两个阶段，首先是对每一层神经网络非监督地预训练该层模型参数，得到各层的抽象特征表示，进而将预先训练好的各层神经网络模型叠加，以构成深度模型，并依据训练数据中的标注信息对整个模型的参数进行调优（finetune），以提高算法的准确性，从而体现出深度神经网络复杂模型带来的表示能力提高的优势。随着深度学习技术的自身发展[5,6]，深度模型逐渐不再依赖非监督的预训练，而是直接学习出结构化的模型并用于预测，特征学习也即通过深度模型训练得到的层次化的抽象特征。

3 文本大数据内容理解

由于语言本身是一个复杂的结构对象，借助于特征学习方法可以较好地刻画语言的复杂结构，从而实现对文本大数据的内容理解。基于特征学习的文本大数据内容理解目前主要从两个方面展开：第一个方面是面向非结构化文本，以词汇为基本单元，抽象词汇的特征，进而组合以表示语句的特征，并在特征表示的基础上实现对文本内容的理解；第二个方面是面向结构化知识数据，以知识表示三元组为基本单元，从非结构化文本中抽取出计算机可操作的结构化知识，实现知识的发现、推理等，从而理解文本的内容。

3.1 面向非结构化文本的内容理解

词汇作为自然语言的最小组成单元，学习其特征是让计算机理解词汇进而理解文本的基础；在理解词汇的基础上，阐述语义组合方法，语义组合通过将词汇组合成短语、语句的特征表示，从而让计算机理解文本大数据内容。

3.1.1 词汇理解

在计算机中，所有的字符都是以固定的编码形式表示，例如，汉字“中”在Unicode编码中表示为“4E2D”，字母“A”的Unicode编码为“0041”。计算机中的文字是由无任何意义的编码拼接而成，均无法直接应用于文本理解。因此，一种能够刻画词汇语义特征的表示方式是实现词汇语义理解的关键。

以词汇为基本单位，旨在研究建立合适的词语表示模型，经典的当属以WordNet[7]和知网（HowNet）[8]为代表的人工编制的知识库。WordNet中每个词项（synsets）都代表词汇的一个具体含义，词项间通过词义的语义关系建立联系，形成完善的词汇网络，以表达词汇语义。知网则是把概念与概念之间的关系以及概念的属性与属性之间的关系构成网状的知识系统，知网定义义原为最小的语义概念单元，并通过义原对义项的结构属性相互关系描述词汇语义。这类人工知识库对词汇的语义描述虽然准确，但是其规模小，缺乏可扩展性和自适应能力，难以满足文本大数据语义理解的需要。

利用特征学习方法实现词汇的语义表示源自神经网络语言模型，语言模型的训练目的是最大化词汇出现的概率分布。在参考文献[9]中，作者基于前向神经网络语言模型，随机初始化训练语料库中的词向量表示，以海量文本作为训练数据，假设在文本中套用滑动窗口产生的短句为正例样本f，将滑动窗口中的某个词随机替换为词典中的任一词所产生的错误短句为负样本f'，并令正例样本的得分比负例样本的得分高，以Hinge loss为目标函数，该目标函数在正例样本和负例样本中划分距离为1的边界，从而利用反向传播算法更新词向量，通过训练得到的词汇表示向量，使得相似的词的特征表示也相似。

由于神经语言模型复杂，基于多层神经网络结构计算词汇表示向量，存在计算量较大的问题，训练时间往往需要几天甚至数周。Mikolov等人[10]提出了Word2vec模型，该模型极大简化了多层神经网络结构，仅包含一层投影层，使得计算效率大幅提高。该模型包括连续词袋模型（continue bag of words，CBOW）和Skip-gram模型两种词向量的训练方法。CBOW模型的目标是给定窗口为n的上下文w_c，预测中间的词w_i，其中，投影层为对所有的上下文词向量求平均值，即 $h=\frac{1}{n}\sum_{c=1}^{n}w_c$ ，并利用投影层预测目标词w_i的概率；Skip-gram模型的目标则是给定目标词w_i，预测上下文的词w_c的概率。

3.1.2 语义组合

词汇特征表示向量在一定程度上解决了词汇的语义理解问题，语义组合则是将词汇组合成词组或者语句的语义表示形式，已实现语句级的语义理解。语义组合符合人们理解语句的方式，人们理解语句不是通过直接记忆句子，而是在理解词语和词语组合方式的基础上理解句子的含义。语义组合的目的是将基本的词语单元组合，以表达复杂语句的语义，语句整体的语义看作部分语义的组合函数。因此，语义组合是词汇语义理解向语句语义理解的重要手段。语义组合函数定义为[12]：$p=f(u, v, R, K)$，其中，u、v表示待组合部分，R表示u、v间的关系，K表示用于语义组合的其他上下文知识。

若将R定义为简单的线性关系，则可以实现基于加法$p=u+v$和乘法$p=u\cdot v$的组合函数，这种组合方式虽然简单，但在组合时忽略了词在文本中出现的顺序，即$u+v=v+u$或$u\cdot v=v\cdot u$，存在明显的缺陷。这导致不同

含义的词组可能有相同的表示形式，例如“种子植物”和“植物种子”有相同的表示，但是这两个词组前者描述一类植物，后者表示种子，意义并不相同。有研究显示，英文文章的含义 80%来自于词的选择，20%来自于词的顺序，因此忽略词序对语义理解有较大的损失。

基于特征学习的复杂模型由于符合语义组合的方式、刻画语句的特征，获得广泛的关注，并在语句的语义理解上取得很好的效果。递归自编码（recursive autoencoders）[13]是一种非线性的语义组合模型，它以递归的方式组合自编码网络，构建短语或句子的语义特征表示。递归自编码模型是由自编码模型组合而成，自编码模型是一种非监督的神经网络模型，该模型以输出数据约等于输入数据为训练目标更新模型参数，得到编码后的隐藏层g为模型输入数据的特征。如图2所示，u、v为待组合的两个词语的特征表示向量，利用自编码模型计算组合后词组的特征表示 $p=f(W[u:v]+b)$ ，为了训练词组的特征表示，模型解码词组特征得到 $[u':v']=f(W'p+b')$ ，并以 $[u:v]\approx[u':v']$ 为目标训练模型的参数和词组的特征表示向量。如图3所示，在得到二元词组的语义组合表示的基础上，可以递归地扩展为一棵二叉树的结构以实现语义扩展。目前，将句

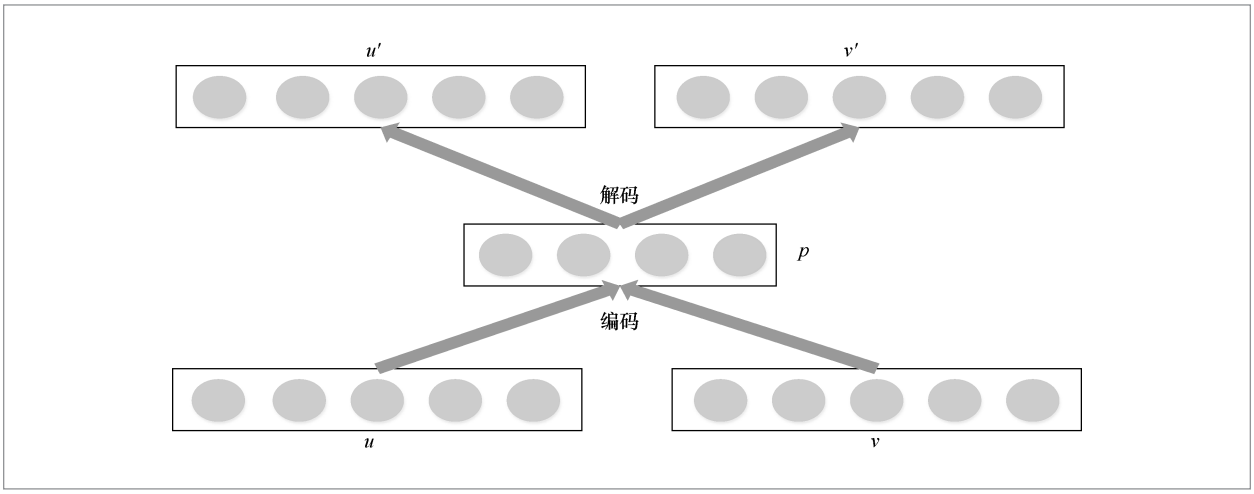


图 2 自编码模型

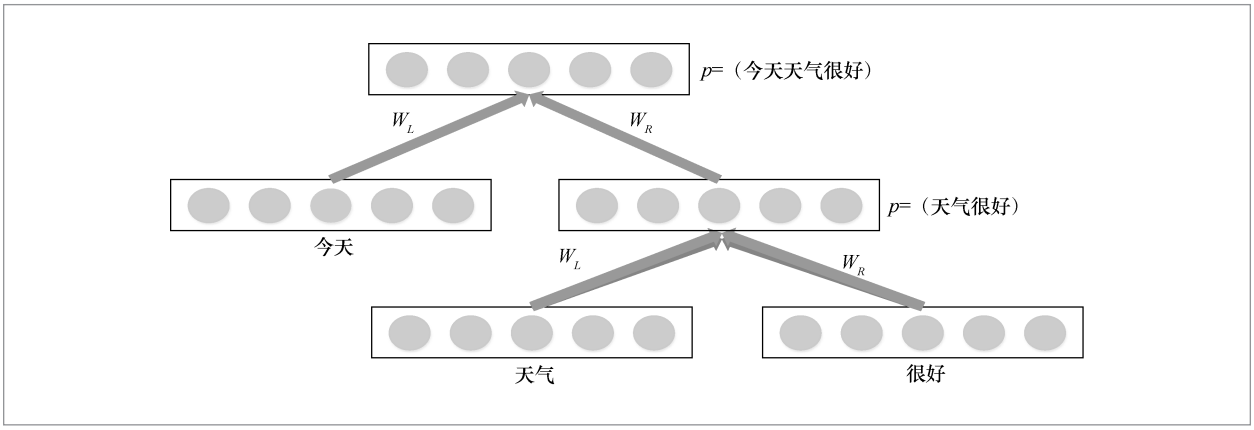


图 3 基于递归自编码的语义组合模型

子构建成树有两种方式：一是利用贪心算法构建一棵树，对于长度为n的句子，计算$n-1$个连续二元词组所构造的自编码模型错误率，选择错误率最低的两个节点组合构成一棵二叉树，在剩下的$n-2$个节点中，继续选择自编码模型错误率最低的连续两个节点组合，直到组合至根节点为止；二是利用语法树构建递归自编码，该方法将句子解析为语法树的形式，这不仅降低了计算复杂性，还保留了句子的语法结构，因此语义组合后能得到更好的语句特征。

利用语义组合方法构建的抽象语句特征表示，可以更好地识别出语句的内在语义，使得相似的语句有相似的特征表示，从而用于语句的情感挖掘、词组相似性、同义语句识别等语义理解任务中。

3.2 面向结构化知识的内容理解

结构化知识是文本内容理解的产物，同时也可用于文本的内容理解。知识数据作为搜索引擎、智能问答重要的信息源，扮演着越来越重要的角色。通过知识图谱可以建立实体的关系网络，赋予丰富的语义信息，从而为文本理解提供基础。

3.2.1 知识表示

将知识表示成计算机可计算的符号化形式，是让计算机理解知识的基础。对知识表示的研究伴随着计算机的整个发展阶段，提出了一系列表示方法，如谓词逻辑表示方法、框架式表示方法、产生式表示方法和面向对象表示方法等，不同的知识表示方法对问题解决有不同的帮助。良好的知识表示方法应能满足不同类型使用者的要求，一般来说，对知识表示的要求应考虑以下几个方面：表示知识的范围要广泛，表示的形式要适合于推理，并且要具有可解释的能力。

随着语义网的发展，将知识以本体（ontology）的形式进行组织，以描述概念和概念间的关系，这已经成为重要的知识表示方式；但是，由于本体的结构过于复杂，近年来语义结构相对简单的知识图谱成为知识表示的热门发展方向。

通常，知识图谱包括大量的实体（如奥巴马、夏威夷）、实体的语义类别（如奥巴马属于政治家分类，夏威夷属于城市的分类）和实体间的关系（如奥巴马和夏威夷的关系是奥巴马出生于夏威夷），并以三元组的形式表示（主体，关系，客体），记作(e_l, r, e_r)（如（奥巴马，出生于，夏威夷））。

由于知识图谱的重要作用，学术界和工业界都在努力构建大规模知识图谱，以满足实际应用需要，其中，典型的知识图谱包括Freebase、NELL（never-ending language learning）等。Freebase是以众包的形式构建的知识图谱，因而包含一定的噪音数据，目前已包含大于4 000万个实体、大于 20 000种关系，共大约 19 亿条记录；而NELL项目自2010年以来，不断地从互联网中抽取结构化数据，且不停地迭代更新已有数据的置信度，目前已累计超过5 000万条知识数据。

3.2.2 知识发现

利用特征学习表示知识数据是在词汇特征表示捕捉词汇语义的基础上，构建关系的表示方法，进而实现结构化知识的发现。其中，经典的工作是TransE模型[14]，该模型将三元组中的关系看作主体到客体的翻译，使得三元组满足$e_l + r \approx e_r$的线性转换。利用特征表示向量描述实体和关系，可以更加容易地计算实体间的语义关系。但是该方法不能很好地刻画多对一、一对多或多对多的关系，例如在多对一的关系中，在关系r和客体e_l^r的特征表示向量相同的情况下，由于三元组满足$e_l \approx e_r - r$的

映射要求，使得不同主体的特征表示也会相同，这显然不符合特征的语义表示，因此该模型未来还有继续改进的空间。

在得到实体、关系的特征语义表示的基础上，可以进一步实现关系的抽取和发现。例如，给定主体e_l和客体e_r，通过判断与e_r-e_l最相似的关系特征表示向量r，确认两个实体间的关系；或在给定主体e_l和关系r的情况下，判断与e_l+r最相似的客体e_r，从而发现新的三元组知识数据。实验显示，通过简单的向量加减法可以发现新的事实数据或判断实体间的关系，这极大提高了知识发现的效率。

3.2.3 知识推理

计算机的推理能力是计算机智能的重要体现。在知识图谱中，基于实体关系的推理是发现隐藏知识的重要手段。传统的基于规则的推理方式，由于完全依赖人为定义，发现的关系受限于人为定义的规则库，因此自动化的关系推理是丰富现有知识图谱的重要手段。在基于线性关系发现知识数据的基础上，进一步扩展线性转换关系，可以实现多关系组合推理[15]，给定两个相关联三元组(e_l,r,e_r)和(e_r,r',e_r')，根据三元组的线性变换规则，可以认为在实体、关系的特征语义空间中，多个三元组间存在$e_l+r+r'\approx e_r'$的组合推理关系，从而实现知识的推理。

3.2.4 隐式关系发现

知识图谱是对文本大数据内容理解的产物，同时，知识图谱作为丰富的知识资源可以反作用于文本的内容理解。由于个体文档通常只包含少量的关系数据，这些关系数据可能不足以体现完整的实体关系网，但是通过与已有的知识图谱匹配，可以完善实体间的关系，从而发现现有文本中无法挖掘的隐式关系，满足文本数据深层次内容理解的需要。

4 基于特征学习的文本内容理解发展趋势

基于特征学习的方法在文本内容理解问题上已经取得了一系列突破，未来结合网络大数据的涌现，对文本内容理解的研究还将继续发展。针对非结构化文本的内容理解，深度学习由于其可以抽象高层次的概念特征，是未来重要的研究方向；而针对结构化知识的内容理解，知识图谱可以结构化、形式化地刻画文本的语义内容，进而实现关联推理，是实现文本内容深度理解的重要手段。

深度学习作为非结构化文本内容理解的重要方法，未来将继续探索适合文本内容理解的模型，以提高内容理解的准确性。语言是一种序列模型，语言本身具有一定连续性，因此一个能刻画语言时序特征的模型是实现文本内容理解的重要基础。由于递归神经网络具有一定的时序性和记忆性，利用递归神经网络训练文本的特征符合语言的形式，在机器翻译、自然语言生成等应用中都取得较好的结果，递归神经网络正逐渐成为文本内容理解的重要模型。在递归神经网络模型的基础上，有研究进一步提出有长期记忆能力的递归模型，并将该模型用于自动问答中[16]，取得了较好的结果。具有较强记忆能力的模型对于文本内容理解起着重要的作用。

同时，对基于深度学习方法自动学习的文本特征的可解释性也是未来研究的方向。不同于直观的人工定义特征，通过特征学习方法得到文本抽象特征，其对人而言的可解释性并不强。最近，Google的研究人员提出了Deep Dream方法，可视化

地针对图像识别的深度模型各层特征[2]。对于文本而言，目前对于文本的抽象特征以及模型自身的可解释性都还有待进一步研究，只有理解了模型及其抽象特征，才能更好地实现文本内容的理解。

知识图谱作为结构化知识的重要组织形式，刻画实体关系的演化是重要发展方向。实体间的关系是不断演化发展的，具有时序性，因此有其自身的生命周期，绘制一张动态的知识关系网，对文本大数据内容理解的实时性有很大帮助。此外，目前的知识图谱围绕实体展开，描述实体间的关系；未来如何从文本大数据中抽取事件信息，实现事件的发现和推理，是文本大数据全面深入内容理解的重要方向。

2 http://googleresearch.blogspot.ch/2015/06/inceptionism-going-deeper-into-neural.html

5 结束语

随着文本大数据的涌现，文本处理已经从数据不足转向数据过量，虽然文本大数据主要是无标注或者弱标注的数据，但是这类数据正好为特征学习方法提供了数据基础，进而实现了特征发现基础上的文本语义理解。基于特征学习的文本内容理解有了许多探索和突破，但是由于自然语言自身的复杂性、模糊性，特征学习需要更为准确的结构以刻画自然语言。相信随着特征学习技术的发展和对自然语言本身认识的加深，对文本大数据的内容理解能力一定会进一步提高。

参考文献

[1] Bengio Y. Deep learning: theoretical motivations. Presented at the Deep Learning Summer School, 2015

[2] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798~1828

[3] Bengio Y. Learning deep architectures for AI. Foundations and Trends in Machine Learning, 2009, 2(1): 1~127

[4] Hinton G E, Osindero S. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7):1527~1554

[5] Srivastava N, Hinton G, Krizhevsky A, *et al*. Dropout: asimple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 2014, 15(1): 1929~1958

[6] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 2010: 807~814

[7] Miller G A. WordNet: a lexical database for English. Communications of the ACM, 1995, 38(11): 39~41

[8] 董振东，董强，郝长伶. 知网的理论发现. 中文信息学报，2007, 21(4): 3~9
Dong Z D, Dong Q, Hao C L. Theoretical findings of HowNet. Journal of Chinese Information Proceeding, 2007, 21(4): 3~9

[9] Collobert R, Weston J. A unified architecture for natural language processing : deep neural networks with multitask learning. Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 2008

[10] Mikolov T, Corrado G, Chen K, *et al*. Efficient estimation of word representations in vector space. Proceedings of Workshop at ICLR, Florida, USA, 2013: 1~12

[11] Maaten L V D, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008(9): 2579~2605

[12] Mitchell J, Lapata M. Composition in distributional models of semantics. Cognitive Science, 2010, 34(8): 1388~1429

[13] Socheer R, Perelygin A, Wu J Y, *et al*.

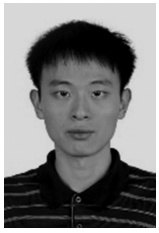
Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), Washington DC, USA, 2013: 1631~1642

[14] Bordes A, Usunier N, Garcia-Duran A, *et al*. Translating embeddings for modeling multi-relational data. Proceedings of Conference on Advances in Neural Information Processing Systems (NIPS), South Lake Tahoe, Nevoda, US, 2013: 2787~2795

[15] Garcia-Durran A, Bordes A, Usunier N. Composing relationships with translations. Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), Lisbon, Portugal, 2015: 286~290

[16] Sukhbaatar S, Szlam A, Weston J, *et al*. End-to-end memory networks. arXiv Preprint arXiv:1503.08895, 2015

作者简介



袁书寒，男，同济大学博士生，主要研究方向为自然语言处理、深度学习、大数据分析。

向阳，男，同济大学教授、博士生导师，主要研究方向为大数据分析、云计算、语义计算、管理信息系统，主持和参与多项国家“973”计划、“863”计划、国家科技支撑计划、国家自然科学基金项目，近年来发表论文50余篇。

鄂世嘉，男，同济大学博士生，CCF学生会员，主要研究方向为云计算、知识图谱、大数据系统。

收稿日期： 2015-08-26

基金项目： 国家重点基础研究发展计划（“973”计划）基金资助项目(No.2014CB340404)，上海市科委科研计划项目(No.14511108002)

Foundation Items: The National Basic Research Program of China(973 Program)(No.2014CB340402), The Science and Technology Planning Project of Shanghai (No.14511108002)

论文引用格式： 袁书寒，向阳，鄂世嘉. 基于特征学习的文本大数据内容理解及其发展趋势. 大数据，2015030

Yuan S H, Xiang Y, E S J. Text big data content understanding and development trend based on feature learning. Big Data Research, 2015030