软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2022,33(1):102−128 [doi: 10.13328/j.cnki.jos.006304]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

自然语言处理中的文本表示研究*

赵京胜[1,2], 宋梦雪[1], 高 祥[1], 朱巧明[2]

[1](青岛理工大学 信息与控制工程学院, 山东 青岛 266520)

[2](苏州大学 计算机科学与技术学院, 江苏 苏州 215021)

通信作者: 赵京胜, E-mail: zhao5199@163.com

摘 要: 自然语言处理是人工智能的核心技术, 文本表示是自然语言处理的基础性和必要性工作, 影响甚至决定着自然语言处理系统的质量和性能. 探讨了文本表示的基本原理、自然语言的形式化、语言模型以及文本表示的内涵和外延. 宏观上分析了文本表示的技术分类, 对主流技术和方法, 包括基于向量空间、基于主题模型、基于图、基于神经网络、基于表示学习的文本表示, 进行了分析、归纳和总结, 对基于事件、基于语义和基于知识的文本表示也进行了介绍. 对文本表示技术的发展趋势和方向进行了预测和进一步讨论. 以神经网络为基础的深度学习以及表示学习在文本表示中将发挥重要作用, 预训练加调优的策略将逐渐成为主流, 文本表示需要具体问题具体分析, 技术和应用融合是推动力.

关键词: 自然语言处理; 文本表示; 向量空间模型; 主题模型; 图模型; 深度学习; 表示学习

中图法分类号: TP391

中文引用格式: 赵京胜, 宋梦雪, 高祥, 朱巧明. 自然语言处理中的文本表示研究. 软件学报, 2022, 33(1): 102–128. http://www.jos.org.cn/1000-9825/6304.htm

英文引用格式: Zhao JS, Song MX, Gao X, Zhu QM. Research on Text Representation in Natural Language Processing. Ruan Jian Xue Bao/Journal of Software, 2022, 33(1): 102–128 (in Chinese). http://www.jos.org.cn/1000-9825/6304.htm

Research on Text Representation in Natural Language Processing

ZHAO Jing-Sheng[1,2], SONG Meng-Xue[1], GAO Xiang[1], ZHU Qiao-Ming[2]

[1](School of Information and Control Engineering, Qingdao University of Technology, Qingdao 266520, China)

[2](School of Computer Science and Technology, Soochow University, Suzhou 215021, China)

Abstract: Natural language processing is the core technology of artificial intelligence. Text representation is the basic and necessary work of natural language processing, which affects or even determines the quality and performance of natural language processing systems. This study discusses the basic principle of text representation, the formalization of natural language, the language model, and the connotation and extension of text representation. The technical classification of text representation on a macro level is analyzed. The mainstreams of text representation technologies and methods are analyzed, induced and summarized, including vector space model, topic model, graph-based model, neural network-based model, and representation learning. Event-based, semantic-based, and knowledge-based text representation technologies are also introduced. The development trends and directions of text representation technology are predicted and further discussed. Neural network-based deep learning and representation learning on text will play an important role in natural language processing. The strategy of pre-training and fine-tune optimization will gradually become the mainstream technology. Text representation needs specific analysis according to specific problems. The integration of technology and application is the driving force.

Key words: natural language processing; text representation; vector space model; topic model; graph model; deep learning; representation learning

信息和大数据时代, 海量的数据可以被获取、存储和应用. 文本作为人类交流、沟通和记录信息的工具, 是数据呈现的主要形式. 据统计, 互联网上甚至大多数组织机构中, 约 80% 的信息是以文本的形式存在的. 大量的文本

* 基金项目: 国家自然科学基金 (61773276, 61836007)
收稿时间: 2020-12-01; 修改时间: 2021-01-10; 采用时间: 2021-02-28; jos 在线出版时间: 2021-04-20

数据可以借助计算机进行分析和处理, 从中发现或挖掘有用的模式和知识, 自然语言处理就是完成这种任务的学科领域[1]. 自然语言处理借助计算机对文本进行处理, 处理的对象是真实的人类语言, 采用的方法是计算机算法或模型, 成果是诸如文本分类、文本摘要、机器自动翻译等的实际应用, 以及在此基础上的语言学理论、人脑语言机制等的深入研究. 自然语言处理的最终目标是让机器能准确地理解人类语言, 并自然地与人类进行交互. 在当前和今后很长一段时间内, 自然语言处理领域的研究重点是探索计算机如何表示、存储和处理人类语言, 设计相应的系统实现自然语言处理任务, 并评估这些系统的质量. 这种系统是采用人工智能算法或模型, 编制计算机程序模拟人的自然语言处理机制实现的. 这里有一个核心问题是如何将人类真实的自然语言转化为计算机可以处理的形式, 一般也称为自然语言的形式化或数字化, 在自然语言处理领域通常称为文本表示. 文本表示也称语言表示, 是对人类语言的一种主观性约定或描述, 是认知科学和人工智能领域中的共性和基础性问题. 认知科学认为语言表示是语言在人脑中的表现形式, 影响或决定着人类对语言的理解和产生. 而人工智能认为语言表示是指语言的形式化或数学化描述, 在计算机中表示语言并通过计算机程序自动处理, 比如词向量就是以数值向量的形式来表示一个词. 由此可以看出, 文本表示完成自然语言数据的数字化, 是自然语言后续处理的基础性工作.

自然语言的复杂性、多样性和歧义性使得自然语言处理任务非常困难, 这就要求在应用领域中, 根据处理的要求, 把自然语言以一定的数学形式严密而合理地表示出来, 然后通过机器学习, 实现自然语言的机器处理. 在计算学科, 一般认为数据决定了机器学习的上限, 而算法只是尽可能逼近这个上限, 借助计算机进行自然语言处理所用到的数据就是文本表示的结果. 随着计算机技术的发展和自然语言处理研究的深入, 文本表示越来越受到重视, 大量的文本表示模型被提出并得到应用, 这些方法、框架和工具进一步提升了自然语言处理系统的性能. 但是总体来看, 自然语言处理中的文本表示性能依然低下, 亟待进一步探索和研究更具实用化、通用化的表示方法和模型[2].

本文以文本表示为研究对象, 对文本表示的基本原理、相关概念和主流方法进行探索和分析, 对文本表示技术面临的挑战和未来发展趋势进行了预测. 全文安排如下: 首先分析文本表示的基础; 接下来在对文本表示进行宏观分类的基础上, 归纳和总结主流文本表示的技术和方法, 包括向量空间模型、主题模型、基于图的模型以及基于神经网络的模型等, 对文本表示相关的深度学习和表示学习进行了较全面的探讨. 最后对文本表示未来发展趋势、文本表示与深度学习的关系进行了预测. 以期对文本表示有一个全面的了解, 在实现自然语言处理相关任务时能合理选择文本的表示.

1 文本表示基础

不论什么语种的自然语言文本形式上都是特定符号组成的线性序列, 以中文为例, 基本的符号包括汉字、标点符号以及其他符号 (比如数字、拼音字母、数学运算符号) 等. 从语言学角度看, 这些符号组合成的序列表达了一定语义, 由人根据需要按一定的语法规则组合而成并进行交流、沟通和记载, 是对现实世界的描述, 人脑可以理解和分析, 文本最核心的是其语义信息. 语言基本的机制如图 1 所示[3,4].

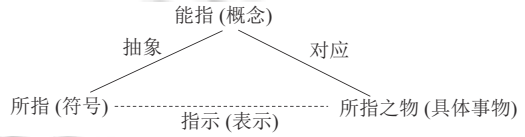

图 1 语言三指

图 1 中所指表示语言的符号, 是符号学意义下的语言构成要素. 不同语种的自然语言有不同的符号体系, 这些符号由人创造并使用, 是一个动态的集合, 比如中文的“桌子”、英文中的“desk”等. 单个或孤立的符号可以组成短语、句子、段落或篇章等复杂的符号集合或符号序列, 这种组合需要相关的语法规则约束. 符号或符号序列本身并不具有语义的要求, 不同语种的语法序列是有差异的, 同一语种的语言符号序列也有索绪尔所分离的“语言”和“言语”的差异[5]. 能指表示人脑对所获取到的语言信息抽象后得到的概念、分类等, 是人脑理解之后的结果, 也可以认为是语言的意义, 即语义. 人脑理解语言的语义, 语义信息在人脑加工处理后可以生成新的语义并转化为新的语言符号序列, 实现人际之间的交流和沟通. 所指之物比较简单, 表示现实世界的实际物体[6].

尽管认知科学、脑神经科学和心理学的研究对人脑语言机制中的语言信息的存储和处理研究有许多突破, 但

其核心原理还不是特别清楚. 用计算机模拟人脑的语言处理机制进行自然语言处理尚缺乏强有力的理论支撑, 再加上自然语言的歧义性、复杂性和多样性, 以及计算机本身算力的有限性和相关算法、模型的局限性, 使得自然语言处理尽管应用前景广泛, 但难度很大, 具有挑战性[7,8].

自然语言处理领域的研究伴随着计算机, 特别是人工智能的产生而产生, 并随着相关技术的发展而发展, 主要经历了 5 个阶段[9−13]: (1) 20 世纪 50 年代之前的萌芽时期. 计算机的诞生、人工智能的提出为自然语言处理提供了基础和动力, 催生了人类借助计算机探索语言机制和语言应用的自然语言处理的研究和应用; (2) 20 世纪 60–70 年代是自然语言处理的快速发展时期. 以乔姆斯基的形式语言理论为代表的符号主义和生成语法, 通过人工的方式, 利用语言知识归纳形成语言规则, 建立语言知识库, 进而实现语言的表示、推理和应用; (3) 20 世纪 70–80 年代是自然语言处理研究的低谷时期. 由于前期对计算机处理自然语言的难度估计过于乐观, 许多系统并未达到预期目标, 引起了业界对自然语言处理, 甚至是人工智能的怀疑; (4) 20 世纪 90 年代–2010s, 自然语言处理进入统计时代. 借助概率统计等数学工具、信息论以及机器学习方法实现自然语言处理的统计和概率分析, 大量的基于统计学习的理论、方法和模型被提出并应用于自然语言处理实际领域, 取得了许多实用化的成果; (5) 2012 年以来, 基于深度学习的自然语言处理成为主流方法. 建立在神经网络基础上的深度学习在语音、图像等领域取得了良好的效果, 自然语言处理领域引入深度学习, 在许多任务上取得了明显的质量提升.

近年来, 自然语言处理进入融合发展期, 基于统计、基于神经网络和深度学习以及基于规则方法的自然语言处理互相融合、互相促进. 不论哪个阶段, 自然语言处理应用的基础性工作就是语言的形式化或数字化, 也就是将符号的文本转化成数字化的文本表示, 比如向量、矩阵、张量等, 然后借助计算机的高速度和大容量存储实现后续的处理和应用.

1.1 什么是文本表示

自然语言处理是语言学、数学和计算机科学的交叉学科. 从语言学的角度看, 语言是一个符号系统, 包括内容和形式. 语言的内容是对现实的抽象、是人与人之间信息的表达和交流, 语言需要通过形式来反映内容, 相同的内容可以通过不同的语言形式来表达, 这既是一种语言中有限符号的无限表达的基础, 也是不同语言之间相互翻译的基础. 基于语言学的文本表示在分析文本基本组成要素的基础上, 重点分析文本成分之间的组成和结构, 实质上是对语法的分析和表示, 形成词法和语法规则. 在自然语言处理的初期, 这种基于规则的文本表示是主流方法. 从数学的角度看, 数学几乎是所有自然科学, 甚至是许多人文学科的基础. 借助数学模型对自然语言进行建模分析, 既是自然语言形式化、数字化的必然趋势, 也是一项基础性工作. 通常利用集合、向量、矩阵等进行文本表示, 引入概率与统计、函数、图论、网络技术、信息论等进行文本分析. 从计算学科的角度看, 利用计算机对自然语言文本进行处理是人工智能的重要分支, 也被称作计算语言学 (computational linguistics, CL) 或人类语言技术 (human language technology, HLT). 自然语言处理的核心目标是让计算机处理人的语言, 而计算机处理符号化的自然语言首先要解决语言符号的输入和存储, 然后通过统计计算、机器学习等技术实现文本分析、处理和生成等后续操作.

人-机、机-机之间的语言信息交流建立在不同层面的人工智能基础上, 如语音识别、文本挖掘、信息抽取、文本理解和生成等. 这些技术首先需要抽象语言中和意义最为相关的要素, 这就需要对语言进行准确的形式化表示. 自然语言的形式化是指先建立一个符号系统, 确立符号连接成合法序列的规则, 约定合法符号串如何表示自然语言中的语义, 定义这些符号可以进行的运算[14,15]. 自然语言的形式化数据模型包括: 数学中的向量、矩阵、集合等; 计算学科中的字符串、链表、树、图等; 生物学中的神经元、神经网络等; 逻辑学中的命题、谓词、产生式、推理规则等; 物理学中的场、波、谱、价等; 心理学中焦点、模板、意象图式等. 自然语言的形式分析理论包括: 链语法、格语法、范畴语法、转换生成语法、功能合一语法、依存语法、配价语法、词汇语法、优选语义学、Montague 语法、定子句语法、系统功能语法、修辞结构理论、概率语法等. 自然语言的形式化算法包括: 自动机、隐马尔可夫模型、基于规则的线性语法处理算法、基于合一的算法、基于概率的算法、动态规划算法、分布式并行算法等.

在自然语言处理的不同时期, 各种语言形式化理论和方法在词汇、句法、篇章和语义层面获得了一定应用,

有些具有较好的语言学解释. 伴随着互联网的普及和大数据时代的到来, 电子化的自然语言文本或文本的电子化非常广泛, 源文本的获取比较简单, 借助计算机进行文本处理成为必然. 典型的自然语言处理一般流程如图 2.

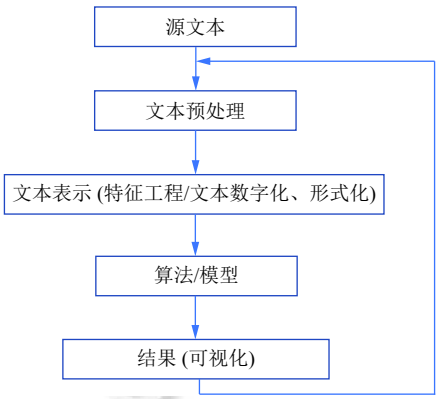

图 2 自然语言处理流程

文本预处理包含文本的清洗、纠错、格式处理, 以及为文本深入分析而进行的分句、分词、词性标注和停用词过滤等. 根据任务的不同, 选择进行相关的预处理, 比如中文分词, 一般的文本处理需要做文本分词, 但有些基于字的文本处理就不需要分词了. 文本预处理中的大多数工作是自动化处理, 有些是人工处理或两者的结合. 算法或模型以及结果的输出是根据具体的自然语言处理任务确定的. 文本表示, 早期的自然语言处理也称为特征工程 (feature engineering), 在整个的流程中处于中间位置, 起着承上启下的作用, 在文本预处理的基础上, 将字符序列的文本转化成计算机可以分析的数字化形式. 文本表示实质上就是一种建模, 涉及到模型相关的数据结构、存储和算法, 主要包括两个步骤: (1) 选择或构建文本表示模型, 也就是要确定用什么样的要素表示非结构化的文本; (2) 文本要素的数值化, 进而实现文本的数值化.

与文本表示密切相关的一个概念是语言模型 (language model, LM), 语言模型是用来判断文本数据合理性的一种机制, 也就是衡量或量化句子的合法性. 语言模型可以根据上下文预测下一个语言单位是什么, 可以从大规模的文本中学习到语义. 自然语言处理在 20 世纪 80 年代以前采用基于规则的语言模型, 也称文法模型[16−18]. 类似于编程语言的处理机制, 由专家归纳整理出自然语言的语法规则, 计算机利用这些规则解析语言文本, 判断其合法性, 或根据规则生成合法的文本. 基于规则的语言模型的解释性很强, 但自然语言的多样性、歧义性和动态性, 以及专家知识的有限性使得人们很难构造复杂的语法和语义组合关系规则, 这就使得基于规则的语言模型的持续性和移植性较差.

20 世纪 80 年代末至 2010s 年, 从统计角度建模的统计语言模型 (statistical language model, SLM) 成为主流方法[19,20]. 自然语言有很多随机因素, 文本序列中的语言成分 (比如词) 的使用可以认为是一个随机事件, 一个文本序列的概率大小可以用来评估其符合自然语言规则的程度. 基于此, 统计语言模型把语言成分序列看作是一个随机事件, 通过概率确定其语言合法性. 例如给定序列中语言成分的样本空间, 也就是词表 (vocabulary) V, 由 V 中 n 个词 $W_1,W_2,\cdots,W_n$ 组成的句子 S, 其概率 $P(S)=P(W_1,W_2,\cdots,W_n)$, 表示 S 存在的可能性, 计算机从大规模语料中学习统计语言模型的参数. 统计语言模型中常采用 n 元语法模型 (n-gram) 进行计算, 当 n=1 时, 称为一元语言模型 (unigram), 序列中每个词都和其他词独立, 也就是和它的上下文无关, 当 n=2 时, 称为二元语言模型 (bigram), 当 n=3 时, 称为三元语言模型 (trigram). 由于语料丰富, 统计语言模型简单、易于实现, 计算结果主要是上下文的条件概率, 借助平滑技术解决数据稀疏问题, 效果较好, 可解释性强, 经过多年的发展, 有大量的计算工具, 因此获得了广泛的应用[21,22]. 但统计语言模型也有比较明显的缺点, 包括注重文本匹配而忽略语义和推理、样本依赖致使非词表 (out of vocabulary, OOV) 问题的处理能力差、无法存储较多的语义信息使得模型失去了泛化能力等[9].

2003 年, Bengio 等人[23]提出了神经网络语言模型 (neural network language model, NNLM), 使用低维、稠密的实值向量表示语言中的组成要素. 比如单词, 利用单词在语料中的上下文进行该单词的分布式表示, 形成词向量. NNLM 解决了基于规则的语言模型的主观性和基于统计的语言模型的高维、稀疏性, 同时通过词向量可获取词之间的相似性, 进而为语言的语义解析提供了基础. 2010 年, 最早的基于循环神经网络 (recurrent neural network, RNN) 的语言模型 RNNLM 由 Mikolov 等人[24]提出, 使语言模型的性能有了较大的提升. 2012 年, Sundermeyer 等人[25]提出了基于长短期记忆循环神经网络 (long-short term memory & recurrent neural network) 的语言模型 LSTM-RNNLM 进一步改善了语言建模的性能. 另外, 基于卷积神经网络 (convolutional neural networks, CNN) 的语言模

型 CNNLM 也取得了成功[26]. 近年来, 基于神经网络的语言模型成为文本表示的主流方法, 特别是深度机器学习解决了传统语言模型的单向性和浅层性, 建立在转换和预训练基础上的深度神经网络语言模型借鉴了语音和图像的底层抽象和迭代思想, 为自然语言处理提供了架构式和普适性的文本表示和应用. 基于深度学习和神经网络, 文本表示采用表示学习机制, 再针对后续应用进行微调[27].

认知语言模型以认知科学为基础, 认知科学是建立在体验基础上的以“现实-认知-语言”为核心原则的交叉学科, 涉及感觉、知觉、意向图式以及范畴、概念、意义等过程. 直观上分析人对语言的理解: 当接受一句话时, 大脑中就会对所描述的客观世界产生心理映射, 形成“内部语言”, 也就是人处理语言的过程是外部语言转化为内部语言的过程; 人脑对内部语言加工处理后, 内部语言再转化成外部语言[28−30]. 类脑语言信息处理的系统采用类似的处理过程, 先确立内部表示的形式, 然后将获取的语言成分转换成计算机可以理解的内部表示, 根据算法或模型处理后, 内部表示转化为外部形式输出. 认知语言学视野下的文本表示主要有两种: 一种是建立在物理符号系统假设基础上的符号主义方式, 将语言的内部表示转化为符号和规则; 另一种是建立在脑科学和神经科学基础之上的联结主义方式, 将语言的内部表示转化为多粒度、多层次神经元的神经网络方式[31,32].

文本表示是自然语言处理的基础性和决定性工作. 文本不同于声音、图像和视频, 文本是对人类思维的较高级抽象, 一般需要背景知识, 文本中的符号除了形式化地推演形成“言内之意”外, 还需要“言外之意”, 包括上下文、语境信息, 甚至是背景或常识知识. 文本表示至少需满足两个条件: (1) 表达效果. 文本表示需将源文本信息恰当、完全地表达出来, 也就是表示过程中应保证语义信息的完整性和一致性. 文本表示空间尽可能地包含原文本空间内的信息, 因为一旦在空间映射时丢弃或歪曲了信息, 则在后续的计算中就无法再获取到这些信息了. (2) 表达效率. 源文本转化成文本表示的代价不能太高, 应尽量减小复杂度, 同时文本表示的结果应便于后续文本处理的实现. 文本表示产出的数据质量直接影响到后续模型的表现, 比如在一个知识图谱构建任务中, 可以将文本表示成分词之后词的 TF-IDF (term frequency-inverse document frequency) 向量, 也可以表示成 LDA、TextRank 或他们之间的结合向量形式, 然后利用模型抽取实体和关系. 无论后续模型怎样, 前面的文本向量表示都会直接影响准确率. 从具体的自然语言处理任务来讲, 作为前期基础性工作的文本表示, 首先是在表达效果上, 不能丢失文本必须的语义信息, 同时表达效率尽可能高, 另外还需要兼顾实用性、针对性和可行性.

1.2 为什么进行文本表示

文本的组成要素是符号, 可以划分成不同的粒度, 比如中文文本中的字、词、短语、句子、段落、标点符号及其他符号等. 文本形式上是不同粒度符号的序列, 这种线性结构可以进一步抽象成树状结构或图状结构. 不论是简单的符号串序列, 还是抽象成的树或图数据形式, 计算机是无法直接对这些文本字符串进行处理的, 必须进行数值化或向量化. 文本表示既涉及单个不同文本粒度内容对象的独立表示, 也包括这些要素组合而成的文本表示. 文本表示的基本流程如图 3 所示.

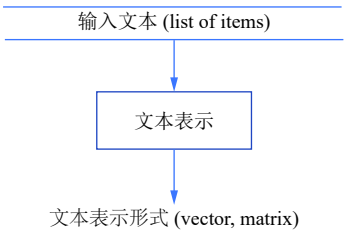

图 3 文本表示基本流程

前面提到, 从认知科学角度理解文本表示, 是大脑对语言信息的存储、加工和处理. 从计算角度理解文本表示, 是存储程序工作原理和编译思想的应用. 从信息与通信角度理解文本表示可以认为是编码 (encoder) 和解码 (decoder) 的实现机制[22,27].

自然语言处理需要进行文本表示, 这是由当前计算机的硬件特点所决定的. 计算机识别和处理的是二进制数据, 任何计算都需要数据的数字化, 自然语言的字符串序列是不能直接通过计算机进行语义相关处理的, 需要借助向量或矩阵等数据结构实现文本的数字化. 同时, 不像语音与图像信息易于实现向量化或矩阵化, 自然语言处理中的文本内容对象难以直接被向量化, 需要研究适合的文本表示模型和方法. 当前, 随着深度学习的飞速发展和广泛应用, 其在图像、视频上的应用效果十分显著, 但对于自然语言处理, 尽管也有一些改进和提升, 但还需要进一步研究和探索. 原因在于语言是人类特有的文明智慧的结晶, 更侧重语义处理而不只是简单的识别, 这对计算机来说是极大的挑战.

简洁、高效的文本表示是自然语言各种处理方法的基础和要求, 也是助力算法效率提升的重要手段和机制. 早期基于规则的自然语言处理系统通过建立知识库和推理机实现文本表示, 以人工分析和处理为主. 传统的机器学习算法采用特征工程的方法, 发现、统计并形成文本中的代表性特征, 包括词频、词性标注、命名实体识别、词干化等文本词法特征, 句法分析、依存关系、位置信息等语法特征, 指代消解、语义角色标注等语义特征, 以及借助 WordNet、HowNet 等知识库形成的知识特征等, 然后实现文本特征数值化, 获得文本表示. 传统特征工程通过人为设计一些准则, 根据这些准则获取原始数据的有效特征, 特征工程一般和最终的预测模型分开进行. 深度学习算法的文本表示直接包含在了神经网络的学习过程中, 利用多层神经网络中各层之间特征的不同抽象和变换实现特征学习, 进而实现文本表示. 经验证明, 良好的文本表示形式可以极大提升后续自然语言处理算法的效果[33,34].

很显然, 自然语言处理需要抽取出文本的特征并实现数字化表示, 特征工程通过人工处理的方式, 一定程度上解决了这个问题. 如果计算机能通过算法自动化地将文本的特征抽取出来, 既解决了人工投入巨大、存在歧义的问题, 又克服了跨领域、多任务的问题, 将极大地提升后续自然语言处理任务的效率. 表示学习作为目前研究和应用的热点领域, 可以自动、有效地获取文本的特征[35], 原因在于: (1) 多粒度语言的统一表示. 自然语言处理中的多粒度语言单位, 包括字、词、短语、句子、段落、文档等的语义表示可以统一到一致的语义空间, 形成低维、稠密的实值向量, 并进而通过两个文本对象之间的几何距离量化这些语言单位之间复杂的语义关系. (2) 多任务处理的有效性. 统一的语言表示空间可以高效执行多任务, 比如针对句子的分词、词性标注、命名实体识别和歧义消解、实体关系抽取、机器翻译等, 统一的语义空间表示使得这些任务的处理更有效、鲁棒性强. (3) 大规模领域数据文本表示的自动获取. 不同领域, 甚至不同语种的自然语言文本, 比如金融、医学、法律等, 传统方法针对领域设计特征抽取算法, 而表示学习可以自动地、针对性地获取文本特征.

归纳起来, 自然语言处理进行文本表示研究的根本原因是计算机不方便直接对文本字符串进行处理, 需要进行数值化或向量化; 另外, 机器学习算法需要数字化的处理, 不仅传统的机器学习算法需要这个过程, 深度学习也需要这个过程; 最后, 不同的文本表示, 对算法效率的影响是不同的, 良好的文本表示形式可以极大的提高算法的效果. 因此, 针对技术发展, 充分利用自然语言处理的硬件支撑能力、借鉴人脑和语言学的研究成果、总结现有方法的优缺点、探索更加有效的文本表示方法是非常有必要的.

2 主流技术和方法

自然语言文本的内容对象可以划分成不同粒度的符号或符号序列, 文本表示可以从两个维度进行考虑: 一个维度针对文本内容对象的不同粒度, 小粒度的如字、词、短语表示, 大粒度的如句子、段落、篇章表示等; 另一个维度按不同的表示形式划分, 分为离散表示 (discrete representation) 和连续表示 (continuous representation). 文本表示也有浅层和深层之分, 这里的浅层文本表示是指提取文本低级别信息作为特征, 比如单个的字或字符, 甚至是中文中的偏旁部首, 忽略文本对象的上下文关系. 深层文本表示提取文本复杂的特征, 一般包含文本内容对象及其所在的上下文之间的关系, 甚至还包含辅助的知识. 尽管文本特征的“浅层”和“深层”不同于机器学习的“浅层”和“深层”, 但文本表示中浅层的特征往往作为深度学习模型的输入, 依靠深度学习模型得到文本的深层表示. 进一步拓展文本表示的分析角度, 有不同的分类方法, 归纳起来, 文本表示的宏观分类如表 1 所示.

自然语言处理中文本的离散表示是一种局部化方法, 将文本抽象成文本内容对象的序列, 根据目标任务确定文本对象, 将这些对象作为离散的特征进行数值化. 特征采取人工、半自动化或自动化方式从文本中获取, 特征的粒度可以是字、词、词组, 甚至是句子、段落等, 特征以文本中的内容和形式为主, 可以引入外部知识或背景知识做辅助, 这种表示几乎可以应用于各种领域、各种结构和各种规模的文本[41,42]. 文本离散表示的代表是词袋模型 (bag of word, BOW), 也称向量空间模型 (vector space model, VSM), 其中的特征可以是不同粒度的文本形式, 比较常用的是词. 文本离散表示最简单的形式是“独热”表示 (one hot representation, one-hot). 词的 one-hot 表示是将词数字化为一个向量, 主要流程包括: 首先确立词典, 词典可以从文本集中产生, 也可从外部导入; 然后将每个词表示成维数为词典长度的向量, 向量中该词对应的位置为 1, 其余位置为 0. 句子或文本的 one-hot 表示也形成一个向量, 向量的维数也是字典的长度, 根据向量中对文本特征数字化的不同, 主要分为两种: 基于布尔表示的形式

(boolean representation) 和基于计数表示的形式 (count-based representation). 前者的元素值是 0 或 1, 数字化过程中对于向量的第 k 个元素, 如果词典中的第 k 个词出现在句子或文本中, 那么其值为 1, 否则为 0; 后者的元素值是非负整数, 数字化过程中对于向量的第 k 个元素, 其取值是词典中的第 k 个词在句子或文本中的出现次数的某种度量值, 简单的取值方法就是统计该词在句子或文本中出现的次数之和, 改进形式包括 TF-IDF、n-gram 等[9]. 文本的离散表示与第 1.1 节的文法语言模型和统计语言模型密切相关.

表 1 文本表示宏观分类

编号	分类依据	类别名称	主要特点	代表性成果与应用
1	表示形式	离散/符号化 连续/分布式	简单、直观; 高维、稀疏 充分利用上下文信息	基于规则、统计, 浅层机器学习模型 基于统计、神经网络, 词向量等
2	计算基础	集合法 代数法 概率论和信息论 图方法 分布式/神经网络	不考虑序列/结构 利用向量、矩阵等进行建模 训练统计模型、依赖数据 考虑结构, 节点、边抽象建模 基于上下文假设理论	信息处理用语言学知识库, 比如词典 Salton、Milokov等[36,37] 传统机器学习模型, 如HMM、SVM等 TextRank、Hits、DeepWalk等 Word2Vec、Sent2Vec、Doc2Vec等
3	文本粒度/规模	字(中文偏旁部首) 词 短语 句子 段落或短文本 篇章或长文本	语言最小单位, 中文研究较多 最小语义单位, 研究重点 词与词之间的语法/语义组合 完整的意义表达结构 文本对象及其结构综合考虑 文本对象及其结构综合考虑	字本位[38,39] one-hot、Word2Vec等 n-gram Sent2Vec Paragraph2Vec、Doc2Vec等 Doc2Vec
4	经典/主流方法	词袋法(BOW/VSM) 基于主题模型 基于图/复杂网络 基于神经网络 基于预训练/调优 跨领域融合方法	关注文本内容, 不考虑词序 文本-主题-词的概率计算 考虑文本对象及其结构 模拟生物特点, 多隐层 迁移学习、多任务学习等 综合运用	Salton[37] LSA、PLSA、LDA等 图理论、复杂网络、知识图谱等 目前的研究与应用热点 ELMO、GPT、BERT等 Text2Image、Image2Text等
5	结构和应用	基于规则 基于统计 基于语义(知识)	专家建立、移植性差 借助统计数学模型 终极目标	Chomskey[3,18] 20世纪90年代自然语言处理各领域 WordNet、HowNet等
6	获取方式	人工定义 自动学习	依赖领域专家 表示学习自动获取	早期主流方法, 目前特定领域/任务 文本表示学习[35,40]
7	信息层次	浅层 深层	简单、效果差 复杂、需要表示学习	基于规则、统计的词法、语法分析 NNLM、DeepNLP等
8	量化表示	特征工程 基于统计 基于语义	人工特征选择或特征组合 自动对文本特征进行统计 终极目标	术语提取、本体构建等 TF、DF、TF-IDF等 语义词典、知识库等

自然语言处理中文本的连续表示也称分布表示 (distributional representation), 基本思想建立在 1954 年 Harris[43]提出的分布假设理论: “上下文相似的词, 其语义也相似”, 以及 1957 年 Firth[44]对该理论的阐述和论证: “词的语义由其上下文确定”基础上. 用一个文本对象附近的其他对象来表示该对象, 是现代自然语言处理的重要创新. 人的语言表达, 重点在语义信息, 一个词及其上下文一起构成精确的语义信息. 换句话讲, 完整语义信息的呈现不应该是离散文本对象的简单叠加, 而应该是文本对象及其上下文语境, 甚至是其他辅助知识共同组成的, 就好像是现实中人以群分, 物以类聚的思想.

分布式表示对文本对象的数值化建模建立在其上下文基础上, 主要由两部分组成[45]: (1) 选择一种方式描述上下文; (2) 选择一种策略刻画文本对象与其上下文之间的关系. 计算形式上以代数法、统计法和语义分析法为主, 可针对所有的文本粒度和文本规模, 获取方式以自动获取为主, 主要分为基于矩阵、基于聚类和基于神经网络的分布式表示 3 种. 基于矩阵的分布式表示最常用的形式是构建一个大小为 $V\times C$ 的共现矩阵 M, 其中 V 是词典大小, C 是所限定的上下文范围大小, 比如在词的表示中, 词 W 的上下文可以为特定窗口中的其他词、W 所在句子或所在文档等. 如果以句子或篇章作为窗口范围, 共现矩阵 M 的每一行就是 W 对应词的向量表示, 每一列是句子

或篇章的向量表示. 对共现矩阵 M 实施数学运算, 比如潜在语义分析模型 (latent semantic analysis, LSA)、潜在狄利克雷分配模型 (latent Dirichlet allocation, LDA) 等, 得到不同的文本表示. 除了构建文本对象及其上下文的共现矩阵, 也可构建文本对象与其上下文之间的语法关系矩阵或语义关系矩阵. 基于聚类的文本表示使用较少, 不再讨论.

基于神经网络的分布式表示 (distributed representation) 不同于建立在代数基础上基于矩阵分解的分布表示, 而是借助于神经网络的非线性变换将文本转化成为稠密、低维、连续的向量, 可以实现不同文本粒度, 比如词、句子、篇章等的表示. 前文 1.1 中提到的神经网络语言模型 NNLM 是最早、最经典的分布式表示模型[23], 在其基础上得到词的表示 Word2Vec[36], 衍生形式包括 GloVe、fastText、ELMO、Transformer、GPT 和 BERT 等. 自然语言处理中近几年提出的 seq2seq 模型、attention 机制, 特别是预训练 (pretraining) 结合调优 (fine-tuning) 技术, 实现了文本表示学习和后续自然语言处理任务的有效结合, 既是对文本表示, 同时也是对具体的自然语言处理任务里程碑式的改进[46], 后文第 2.4 节和第 2.5 节将专门介绍.

文本表示基于规则的方法, 是最早应用于自然语言处理并且与语言学最密切的方法, 具有领域和任务依赖性, 在特定情境下效果优异, 但泛化能力有限, 可以作为主流文本表示的补充或辅助. 文本表示基于统计的方法, 比如 VSM, 建立在文本特征的统计基础上, 简单有效, 但存在语义鸿沟, 有完备的理论基础和较多的支持平台和工具, 目前应用仍较为广泛. 文本表示基于语义的方法, 包括基于主题、基于本体、基于语境框架、基于知识图谱等, 一直以来都是研究的重点. 随着深度学习在自然语言处理中的广泛应用, 文本的分布表示成为主流方法, 特别是表示学习, 可以从文本中自动获取文本的特征, 逐渐成为未来研究和应用的主要形式. 下面对常见的文本表示方法进行分析、归纳和总结.

2.1 向量空间模型

向量空间模型是一种简单有效的文本表示模型, 最早由哈佛大学的 Salton[37]提出, 是一种离散型的文本表示模型. 词的序列形成词袋 BOW, 句子序列形成句子包 (bag of sentence, BOS). 文档 (document) 表示成一个维数很高的向量, 各维对应该文档的特征项 (term), 其数值对应特征项在该文档中的权重 (term weight), 权重值利用具体的自然语言处理任务设定的加权公式进行计算, 体现了特征项在文档中的重要程度, 维数为特征项集合的长度, 该集合也称词典. 这里的文档通常是句子、段落或篇章, 特征项语言单元是根据任务确定的最小语言单位, 不可再分, 可以是字、词、词组或短语等, 特征项可以根据具体任务由人工确定, 也可以借助自动化方式抽取或学习确定. 从集合角度看, 一个文档的内容就是由特征项组成的集合, 表示为: $Doc(t_1,t_2,\cdots,t_n)$, 其中 t_k 是特征项, 特征项的权重反映了特征项在文档中的重要程度. 这样, 文档就可用特征项及其对应的权重来表示, 形成一个向量, 向量的形式为: $Doc((t_1,w_1),(t_2,w_2),\cdots,(t_n,w_n))$, 简记为 $Doc(w_1,w_2,\cdots,w_n)$, 其中, w_k 是特征项 t_k 的权重.

向量空间模型中的数值化权重统计, 如果只关注文档中单个特征项是否包含, 文档的表示变成向量模型的一种特例, 称为布尔模型或 one-hot 模型, 特征的权值只能取二值量 0 或 1, 文档的向量维数是词典的长度. 如果按特征项在文档中出现的频率, 比如词频 (term frequency, TF), 进行统计, 称为基于词频的文档向量空间表示, 向量的权值是特征项在文档中的频率值, 可以是绝对频率, 也可以是归一化后的频率计数, 文档的向量维数是词典的长度. 类似的还有基于词频及逆文档频率 TF-IDF、基于 n-gram 的向量空间表示, 实质上是对文档特征项划分和权重计算方法上的不同, 文档的向量维数是特征项集合的元素个数[47−49].

前面提到过词的 one-hot 表示, 是将词数字化为一个向量, 这种表示方法也称为词向量. 根据任务的不同, 文本中不同的特征项都可以进行类似表示, 对某个特征项, 其向量空间模型表示成一个向量, 维数是特征项的集合长度, 除了对应该特征项的权值为 1 之外, 其余各维的权值为 0. 词或特征项的 one-hot 表示比较简单且容易实现, 但维数过高且难以扩展、词间关系无法体现. 词的表示中除了考虑词本身是否出现和出现次数之外, 还可以考虑词间关系. 根据词间关系, 词的向量表示可以基于共现矩阵的方式构建, 一种方法是基于文档集或文档语料库的词向量, N 个文档组成的文档集, 词在某个文档中出现记为 1, 未出现记为 0, 这样就形成一个 N 维向量. 当然, 对于每维的权值统计也可采用 TF、TF-IDF 或其他计量值. 这种词向量的维数与文档集有关, 也存在维数变化、数据稀疏等问题. 另一种方法是采用特定窗口大小的文本片段, 将其中词的共现作为词向量的构建方法, 维数就是该窗口大小, 每维的权重可以是布尔值 0 或 1, 也可以是其他的量化值, 比如共现的次数.

文本采用向量空间模型表示, 将特征项或其他不同粒度的文本借助向量进行表示, 向量可以组成矩阵, 矩阵通过诸如特征值计算、奇异值分解 (singular value decomposition, SVD) 等矩阵运算可获取优化的文本向量表示[50].

向量空间模型将文档或特征项表示成了一个向量, 模型易于理解、运算实现简单, 有利于后续自然语言处理中相似度、文本分类和文本信息检索等应用. 在早期的自然语言处理, 特别是文本分类相关工作中, 发挥了重要的作用. 但由于特征项的数量庞大, 容易造成数据稀疏和维数灾难. 模型基于特征项之间的独立性假设, 忽略文本中的结构信息, 比如词序、上下文信息等, 再加上特征项的划分和获取没有具体的标准, 以及不关注文本和特征项的意义表示等, 会对模型的泛化和后续应用产生影响. 对 VSM 改进的研究主要集中在两个方面: 一是改进和优化特征项的选择和确定, 获取特征项的语义信息和结构信息, 比如借助图方法[42]、主题方法等进行关键词抽取, 利用深度学习获取词向量, 引入知识辅助, 包括文本外部知识和挖掘内部知识等; 二是改进和优化特征项权重的计算方法, 除了现有的比较成熟的方法之外, 可以在现有方法的基础上进行融合计算或提出新的计算方法.

2.2 基于主题模型的方法

分析作者写文章的过程, 首先确定文章的核心内容, 然后组织材料进行表达, 往往通过划分几个主题来实现, 最后对每个主题选择相应的词汇, 按语法规范将词汇组织起来. 文章分析是这个过程的逆过程, 通过词汇的集合, 确定相应的主题, 最后理解文章的核心思想. 计算机按这种思路进行文本处理符合人的认知过程, 主题模型就是实现这种思想的文本建模形式. 对语料库中的文本, 从词汇出发, 统计学习文本-主题-词之间的关系以实现文本表示, 进而应用于后续的文本处理工作, 比如文本分类、文本摘要等[51−53].

潜在语义分析 LSA 基于分布假设理论和词袋模型, 构建文本的词-文档矩阵 (term-document matrix) X, 每一行为一个词的分布语义表示, 上下文以文档为基本单位, 每一列为一个文档的 BOW 向量表示. 那么, 矩阵运算 $X^{\mathrm{T}}X$ 表示了文档和文档之间的关系, XX^{T} 表示了词与词之间的关系, 对矩阵进行 SVD 分解, 可以得到词和文档的稠密向量表示, 进而发现潜在的主题语义信息.

主题模型通过引入一个“主题 (topic)”作为隐变量, 实现了对 BOW 模型的扩展, 将词和文档之间关联关系抽象为: 文档->主题->词. 主题模型将具有相同主题的词或词组映射到同一维度上, 两个不同的词属于同一主题的判断依据是: 如果两个词有更高的概率同时出现在同一篇文档中, 或给定一个主题, 两个不同的词的产生概率比其他词汇产生的概率高. 主题模型是一种特殊的概率图模型 (probabilistic graphical model, PGM), 数学基础十分完备, 并且基于吉布斯采样的推断简单有效. 假设有 K 个主题 (一般人为设定, 这也是模型可能存在的问题), 就把一篇文章表示成一个 K 维向量, 向量的每一维代表一个主题, 权重代表该文章属于对应主题的概率. 这样, 主题模型计算文本语料库中主题的词分布, 并计算出每篇文章的主题分布.

2003 年, Blei 等人[54]提出的 LDA 是最有代表性的一种主题模型, 通过无监督的 3 层贝叶斯方法发现语料库中有代表性的主题, 并且能够计算出每个文本的主题分布, 并得到每个主题相关的词分布特性. 其基本实现机制如图 4 所示.

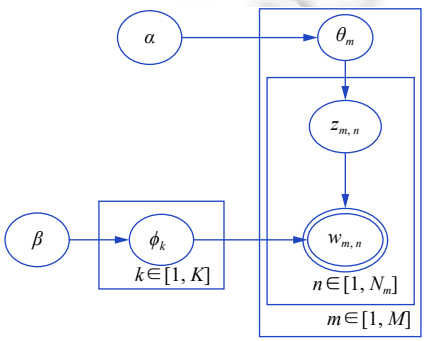

图 4 LDA 主题模型示意图

图 4 中, θ 是文档集关于事先设定的 K 个主题的多项分布, ϕ 是每个主题与词汇表中的词的多项式分布, $z_{m,n}$ 是第 m 篇文本的第 n 个词语 $w_{m,n}$ 的主题, N_m 表示第 m 篇文本中总词数, M 表示数据集中的文本数, α、β 是狄利克雷先验超参数, LDA 的核心思想如式 (1) 和式 (2) 所示:

$$p(w,d)=p(d)\times\sum_{z}p(w|z)p(z|d) \tag{1}$$

$$p(w|d)=p(w|t)\times p(t|d) \tag{2}$$

LDA 涉及到的基本要素是语料库、文档 d 和词 w, 对语料库中的每篇文档, LDA 定义的生成过程主要包括 3 步: (1) 对每篇文档, 从主题分布中抽取一个主题 t; (2) 从被抽到的主题 t 所对应的词分布中抽取一个词 w; (3) 重复上述过程直至遍历文档中的每一个词[55,56].

2.3 基于图的方法

七孔桥问题是 18 世纪著名的古典数学问题之一, 欧拉于 1736 年研究并解决了这个问题, 并由此建立了一个新的学科方向: 图论 (graph theory). 图论将现实问题抽象成一个图, 图直观上是互连节点的集合, 表示为: G=(V,E), 其中, V 表示问题抽象中的节点集合, E 表示边的集合. 图作为一种常用的数据结构, 在计算机中常用邻接矩阵或邻接表进行存储. 现实中的许多问题可以抽象成图进行分析和研究, 比如社交网络、网页链接、交通网络等, 自然语言处理中的文本是否可以表示成图的形式来进行研究呢?

借助于图进行问题建模需解决 3 个基本问题: (1) 表示问题. 获取任务中的节点, 确定如何通过图结构来描述节点之间的关系以及节点和边的属性量化. (2) 学习问题. 图模型中的节点、结构表示学习, 获取相关的参数. (3) 推断问题. 在已知部分信息时, 计算其他信息的分布. 基于图的自然语言表示建模尽管时间不长, 但图结构强大的表示能力得到了充分利用, 降低了数据集构建成本, 融合技术的概率图、神经网络图文本表示为自然语言处理的多个核心任务提供了解决思路, 提高了系统性能, 提供了良好的研究和应用方向[42].

2.3.1 基于经典图论的图文本表示

借助于图进行文本表示最早是由 Schenker 等人[57]提出的, 图中节点是 Web 文本解析后的英文单词, 边是节点间的共现关系, 边的类型是基于 HTML 文件析取所产生的 3 种标签定义, 构建的文本图并未考虑边的权重信息. 以图结构作为文本表示的模型, 较好地保留了文本的结构信息, 简称为 GSM (graph space model) 模型, GSM 构建主要包括 3 个步骤: 第 1 步是获取文本特征, 构建节点集合 V; 第 2 步是定义特征项之间的关系, 确定节点之间的边集合 E; 第 3 步是对节点和边根据需要进行量化, 包括节点属性量化和边的权重量化.

文本特征从原始文本中抽取出来, 可以是字、词、短语、句子或其他形式, 形成节点, 相同的特征项只构造一个节点, 节点的总数就是文本中互不相同的特征项数目, 构成节点集合 V. 以 V 中节点之间的关系构成边, 这种关系最简单的就是共现关系, 若两个特征项出现在一个窗口中, 比如一个句子、特定个数的字符间隔、一个文档等, 窗口内的特征项对应的节点之间就有连边. 文本图可以是有向的也可以是无向的, 所有边的集合构成边集合 E. 节点本身的属性和边的权重根据任务需要进行设定. 除了共现关系形成文本共现图之外, 类似地, 可以构建文本的语法关系图或语义关系图结构[58−60].

2.3.2 基于信息检索的图文本表示

基于图的文本表示最具代表性的应用模型是 TextRank 算法[61], 该算法是由 PageRank 算法[62]改进而来的. PageRank 算法起源于搜索引擎, 最开始用来计算网页的重要性, 基本思想有两点: 一是数量原则, 指向某网页的网页数量越多, 该网页的重要性越大; 二是质量原则, 指向某网页的其他网页的重要性越大, 该网页的重要性就越大. 整个万维网建模为一个有向图, 每个网页为一个节点, 节点之间的链接关系为边, 在此基础上进行迭代计算确定网页的重要性. TextRank 借鉴这种思想, 首先把文本分割成若干特征项并建立图模型, 可以是有向图, 也可以是无向图, 但一般是加权图, 然后进行迭代计算, 对文本中的重要成分进行排序. PageRank 和 TextRank 的迭代过程分别如式 (3) 和式 (4) 所示.

$$s(v_i)=(1-d)+d\times\sum_{j\in In(v_i)}\frac{1}{|Out(v_j)|}s(v_j) \tag{3}$$

$$ws(v_i)=(1-d)+d\times\sum_{j\in Adj(v_i)}\frac{w_{ji}}{\sum_{v_k\in Adj(v_j)}w_{jk}}ws(v_j) \tag{4}$$

其中, $s(v_i)$、$ws(v_i)$ 是第 i 个节点的重要性度量值, d 是阻尼系数, 一般取值为 0.85. 式 (3) 中的 $In(v_i)$ 表示有链接指向网页 i 的网页集合, $Out(v_j)$ 是网页 j 中有链接指向的其他网页的集合, 式 (4) 中的 $Adj(v_i)$ 表示无向图中与 v_i 邻接的节点, 如果是有向图, $Adj(v_i)$ 就是节点 i 的入链 $In(v_i)$, $Adj(v_j)$ 就是 j 节点的出链 $Out(v_j)$. PageRank 算法初始设置每个网页的重要性为 1, 按公式多次迭代达到稳定, 得到的结果就是当前互联网状态下网页的重要性权重值. TextRank 类似进行迭代计算, 简洁有效.

TextRank 算法实现文本表示建模的思想是根据文本要素之间的共现关系构造无向加权图, 主要有两种应用: 一种是用于关键词提取的文本表示建模和算法[63], 首先构建候选关键词图 $G=(V,E)$, 其中 V 为节点集, 一般是在文本预处理后, 将特定词性的词, 如名词、动词、形容词等, 组成候选关键词, 这些候选关键词形成节点集 V, 然后采用共现关系构建节点之间边的集合 E, 两个节点之间存在边当且仅当它们对应的词汇在一个句子或在长度为 K 的窗口中共现; 另一种是用于抽取式的无监督文本摘要方法, 类似地构建图, 但其中节点集合 V 是文本中的句子集合, 将文本分割成句子的表示, 比如句子向量, 所有的句子向量构成图的节点集合 V, 通过句子之间的位置关系, 或通过计算句子向量之间的某种相似性度量确定边的集合 E.

建立在网页链接基础上的典型信息搜索模型除了 PageRank 算法之外, 同时期的 Hits (hyperlink-induced topic search) 算法[64]将互联网网页分成两类: 枢纽或中心页面 (hub page) 和权威页面 (authority page). hub 页面指那些包含了很多指向 authority 页面的链接的网页, authority 页面指那些包含有领域性、主题性等实质性内容的网页. 每个网页赋予两个属性: hub 属性和 authority 属性. HITS 算法的基本思想也包括两点: 一是高质量的 hub 页面会指向许多高质量的 authority 页面; 二是高质量的 authority 页面会被许多高质量的 hub 页面指向. 基于此, 算法采用迭代思想计算每个页面的 hub 值和 authority 值. Hits 算法的迭代过程分别如式 (5) 和式 (6) 所示.

$$a_t(v)=\sum_{(w,v)\in E}h_{t-1}(w) \tag{5}$$

$$h_t(v)=\sum_{(w,v)\in E}a_{t-1}(w) \tag{6}$$

其中, $a(v)$、$h(v)$ 是节点 v 的 authority 值和 hub 值, 下标 t、$t-1$ 为迭代次数, 算法初始设置每个网页的 authority 值和 hub 值为 1, 按式 (5) 和式 (6) 多次迭代到达稳定, 得到的结果就是当前互联网状态下网页的 authority 值和 hub 值. 就如同 TextRank 算法借鉴 PageRank 算法实现文本表示和处理一样, 基于 Hits 算法的文本表示和处理也受到了广泛的关注, 并取得了一些有益的成果, 在此不再赘述.

2.3.3 基于复杂网络的图文本表示

20 世纪末复杂网络研究的兴起, 为文本的图结构建模带来了新的契机. 复杂网络是建立在图论基础上对复杂系统进行建模和分析的理论和方法, 基本定义为: 具有自相似、小世界、自组织、无标度和吸引子中部分或全部性质的网络[65,66]. 复杂网络一般用来描述自然界或人类社会中呈现高度复杂性网络结构的系统, 实质上是由大量节点及其相互作用形成的图结构. 复杂系统中的要素或现象等被抽象为节点, 其间的关系被定义为边. 现实生活中, 包括人际关系网络、电力网络、基因网络、航空网络以及计算机网络等复杂网络形式无处不在, 许多复杂系统都可以通过复杂网络建模进行分析. 因此, 复杂网络不仅是一种数据的表现形式, 也逐渐变成了一种科学研究的手段.

文本复杂网络就是利用复杂网络来描述和建模文本, 研究语言要素及其结构. 通常将文本中的字、词或句子等语言要素表示为节点, 字、词或句子间的关系表示为边, 将文本抽象成图. 从语言学角度看, 人们在表达思想、传递信息时, 往往选择领域或主题相关的语言要素, 但由于时间、场景和语言种类等的不同, 得到的语言表达可能又有较大的差异. 综合来看, 自然语言是一种复杂的、动态的、要素之间相互作用的系统, 语言要素符合幂律特征并具有明显的小世界性. Cancho 和 Sole 最早利用复杂网络来表示文本[67], 采用的方法是基于词语间的共现关系

构建边, 如果两个词语间距离不大于 2, 这两个词语间即可构成边. 与 GSM 相似, 文本中的语言对象如词、短语、句子等定义为复杂网络的节点, 节点间的关系定义为边. 根据边的不同, 常见的文本复杂网络也分为共现关系网络、句法关系网络和语义关系网络等, 也可以是这些不同边构建方法的组合形式[68−71]. 实质上, 文本的复杂网络表示就是一种 GSM 形式, 但是在文本处理的后续任务上, 引入了复杂网络的处理机制, 包括各种中心性度量、小世界性、无标度性以及网络演进的思想等.

2.3.4 基于知识图谱的图文本表示

采用图或网络的形式表示文本, 如果进一步考虑自然语言处理中的知识表示和推理, 或者说在图或网络表示上嵌入知识相关的内容, 就要涉及知识图谱了. 计算机对自然语言的表示经常涉及“语义网络”“认知和语言学知识”“客观世界”等提法, 在实际应用中, 自动问答、机器翻译、推荐系统等既是自然语言处理的主要应用, 也是知识图谱的研究内容. 知识图谱起源于语义网络, 目标是描述现实世界中的各种实体和概念, 以及它们之间的关联关系[72]. 知识图谱直观的表示形式就是一个图 G, 其中节点代表实体, 边代表两个实体之间的关系, G 中的结构定义成三元组形式, 三元组主要有两种: 一种是〈主体 (subject)、谓词 (predicate)、客体 (object)〉, 另一种是〈主体 (subject)、属性 (property)、属性值 (property value)〉.

知识图谱的构建需要自然语言处理技术中的信息抽取, 包括实体抽取和关系抽取, 实体构成 G 中的节点集合 V, 关系构成边的集合 E. 对于文档或文档集合, 提取其中的关键信息, 结构化并最终组织成图谱形式, 形成对文章语义信息的图谱化表示, 比如可以将高频词、关键词、命名实体或主谓宾等语法形式抽取出来进行图谱的组织和表示. 知识图谱借助于图模型, 反映了实体间的结构化关系, 自然语言处理可以借助知识图谱的分析和推理能力完成相关的任务. 以计算机阅读理解为例, 目前的机器实现重点是根据指定文档内容进行推理预测, 而人类阅读理解除了文档内容之外, 更借助文档之外的相关知识进行联想和分析, 利用知识图谱可以在自然语言理解和生成等关键任务中模拟人类的实现机制.

传统的基于经典图论、基于信息检索算法、基于复杂网络和基于知识图谱等的图文本表示需要构建相关的图数据结构, 这些图结构充分利用图通用且强大的表示形式, 建模了不同文本对象以及它们之间的联系. 这些模型都可以归为概率图模型 PGM, 一般借助不同形式的矩阵, 比如图的邻接矩阵 $A \in \mathbb{R}^{N\times N}$ 表示, 其中 N 为构建的图中的节点数. 通常 A 是一个高维、稀疏的矩阵, 直接用 A 表示图数据, 使得建立在其上的机器学习模型效率不高, 因此, 需要实现一种对图数据更加高效的表示方式. 如果从文本中抽取节点、边的信息时实现自动表示学习, 将无监督特征学习和深度学习技术纳入表示学习的范畴, 并充分考虑语言学先验知识和相关的自然语言处理任务, 将大大减少这种数据表示的构建成本和误差[40,73]. 基于表示学习的图文本表示建模建立在深度学习的基础上, 后面再单独分析 (详细见第 2.5.3 节).

2.4 基于神经网络的方法

神经网络原本是一个生物学概念, 人工智能借鉴其实现机制, 通过构造人工神经网络 (artificial neural network, ANN), 由大量信息处理单元, 也称神经元, 互连形成复杂网络结构, 实现对人脑组织结构和信息处理机制的抽象、简化和模拟. 认知和神经科学研究发现, 人的大脑会对感觉器官获取的外界信息进行逐层抽象和提取语义信息. 基于神经网络的深度学习以此为启发, 通过多层网络互联、权值计算捕捉外界输入的组合特征进而提取高层特征, 实现从大量的输入数据中逐级学习数据的有效特征表示. 传统机器学习方法对复杂任务的处理, 往往需要将任务输入和输出人为切割成很多模块或阶段, 逐个分开学习, 比如自然语言理解, 一般需要分句、分词、词性标注、句法分析、语义分析和推理等步骤. 这种方式存在的问题, 一方面每个模块需要单独优化, 优化目标和任务总目标可能不一致, 另一方面模块之间的错误传播会产生逐级放大现象. 基于神经网络的深度学习采用端到端训练 (end to end training) 或端到端学习 (end to end learning), 学习过程中不进行模块或阶段划分, 中间过程不需要人为干预, 统一优化任务总体目标[27,74].

神经网络是一种模型, 不是深度学习, 而深度学习是基于神经网络模型的机器学习方法, 两者在某种角度上看是一致的, 经常作为同义语来用, 不做严格的区分. 多层次的神经网络, 也称深度神经网络 (deep neural network,

DNN) 是深度学习的基础, 本质上是一种特征学习方法, 在语音、图像和视频数据处理中效果明显. 针对自然语言处理, 获取海量无标注文本, 输入多隐层神经网络模型, 经过隐层的非线性变换自动学习文本中的词法、句法和语义特征. 相比传统的浅层机器学习模型, 神经网络模型解决了依赖于人工的复杂的“特征工程”问题, 模型训练得到的词向量可以量化词与词之间的语义关系, 同时模型不需要复杂的平滑算法, 一定程度上解决了数据稀疏带来的计算耗费问题[46].

前面提到 VSM 模型的简化形式 one-hot 模型将词表示成一个向量, 基于图的表示也可以利用词及其上下文构建图得到词的向量表示. 这些方法得到的词向量, 每个维度是词的词典序列或词上下文中的共现、语法或语义的量化, 有具体的含义, 是高维、稀疏的, 可以增添或者删除一些维度, 这种增删是在特定模型上的实现, 只是对计算量有影响. 目前在自然语言处理领域, 词向量是特指嵌入 (embedding) 模型中词的向量表示, 也称词嵌入 (word embedding), 是基于神经网络语言模型 NNLM 或其衍生模型训练得到的低维实数向量. 词的表示是由向量中的所有维度共同决定, 语义分散存储在向量的各个维度中, 训练好后一般不能改动, 单独分析词向量中的一维, 没有具体的含义, 而将每一维组合在一起所形成的向量, 则表示了词的语义信息[75,76]. 神经网络可以对复杂的文本上下文建模, 自动实现词法、句法和语义特征的学习, 训练过程中可根据任务和性能要求进行调参. 基本的 NNLM 实现机制如图 5 所示[23].

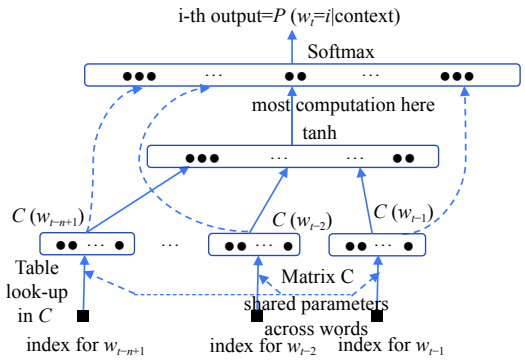

图 5 神经网络语言模型

模型由输入层、隐藏层和输出层组成, 输入第 t–1 到第 t–n+1 个单词的 one-hot 向量表示 w_{t-1}, w_{t-2},…, w_{t-n+1}, 模型预测并输出第 t 个单词的嵌入表示 w_t, 隐藏层通过参数矩阵 $C\in \mathrm{R}^{|V|\times m}$ 将输入的每个词映射为一个向量 $C(i)$, $C(i)\in \mathrm{R}^m$ 表示词典中第 i 个词对应的向量, $|V|$表示单词表中的单词个数, m 表示向量的维度.

词嵌入研究主要涉及两个方面的工作: 一是选取什么特征实现词汇语义的表示; 二是怎样把这些特征有效表示出来. 词嵌入最有代表性的 Word2Vec 模型建立在 NNLM 基础上并对其进行简化处理, 神经网络仅设 3 层, 输入层输入词的 one-hot 向量, 隐藏层不用激活函数, 只是简单的线性处理, 输出层维度跟输入层的维度一样, 采用 Softmax 回归. 模型训练的过程就是通过训练数据获取参数, 主要是隐层的权重矩阵. 根据数据的输入和输出方式, Word2Vec 分为两种: 连续词袋模型 CBOW (continuous bag of words) 是根据目标词上下文中的词对应的词向量, 计算并输出目标词的向量表示; Skip-Gram 模型与 CBOW 模型相反, 是利用目标词的向量表示计算上下文中的词向量. 实践验证 CBOW 适用于小型数据集, 而 Skip-Gram 在大型语料中表现更好. 两者的实现机制如图 6 所示[36].

在训练文本集合中, 以单词 t 为中心, CBOW 方法输入其前 k 个单词和后 k 个单词 (比如 k=2) 的 one-hot 向量表示 w_{t-2}, w_{t-1}, w_{t+1} 和 w_{t+2}, 隐藏层计算目标单词 t 与其前后各 k 个单词的关联概率, 最后输出中心词 t 的嵌入表示 w_t. Skip-Gram 方法输入单词 t 的 one-hot 向量表示 w_t, 隐藏层计算其前 k 个单词和后 k 个单词 (比如 k=2) 与目标单词 t 的关联概率, 最后输出它们的嵌入表示 w_{t-2}, w_{t-1}, w_{t+1} 和 w_{t+2}.

词向量也是基于语言的分布假说理论, 相较于 one-hot 或简单矩阵分解得到的词的向量表示, 以 Word2Vec 为代表的基于神经网络模型训练出来的词向量低维、稠密, 利用了词的上下文信息, 语义信息更加丰富. 根据训练词

向量的思想, 句子、文本等不同粒度的文本都可以类似实现嵌入化表示, 比如 Sentence2Vec、Doc2Vec 等, 甚至是 Everything2Vec. 词向量目前常见的应用包括: (1) 直接使用词向量进行任务处理或使用训练出的词向量作为主要特征扩充现有模型, 如词语相似度计算、语义角色标注等; (2) 词向量作为神经网络中的输入特征, 提升现有系统的性能, 如情感分析、信息抽取、机器翻译等.

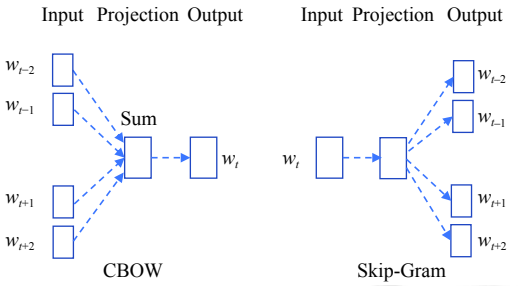

图 6 Word2Vec 词向量实现

Word2Vec 词向量属于典型的基于神经网络概率语言模型 NPLM (neural probabilistic language model), CBOW/Skip-Gram 使用局部上下文窗口 (local context window), 缺乏整体的词和词关系, 负样本采用使得词间关系有缺失. 建立在矩阵分解基础上的 GloVe (global vectors for word representation) 词向量模型[77], 本质上也是基于词共现矩阵, 通过训练神经网络得到词向量. GloVe 利用了全局和局部信息, 使其在训练时收敛更快, 训练周期较 Word2Vec 短且效果更好. 这些预训练词向量的方法是神经网络开始进入语言表示学习 (详细见第 2.5 节) 的主要手段, 推动了基于神经网络的自然语言处理进入了一个新阶段, 提升了系统的性能. 但是仅仅使用目标词特定窗口内的上下文作为目标词的语境词, 缺乏词间相互依赖关系的有效建模, 同时为了避免稀疏, 不考虑词的位置信息, 造成上下文建模方式非常粗糙, 只是词的一种共现统计, 词间相互依赖的关系估计比较粗略, 只能获取到单个词的模糊表示. 再加上简单的上下文词不足以表达丰富的语境信息, 使得基于静态预处理词向量的后续预测任务存在歧义. 但不容置疑的是, 尽管神经网络技术不是新鲜事物, 通过引入神经网络, 给自然语言处理提供了可行的研究思路, 开辟了新的研究方向和应用领域, 为自然语言处理提供了新的契机. 用一个向量代表预训练后的一个词已经成为标准做法, 也是后续基于神经网络的自然语言处理模型的基础. Google 设计的文本表示模型[78], 借鉴了 Word2Vec 的思路, 通过单层神经网络建立词预测模型, 将文档看作是一个特殊的词, 借助分布记忆模型和分布词包模型实现文本表示, 得到文本的向量表示. Facebook 提出的 fastText 文本表示模型[79], 采用了与 Word2Vec 类似的架构, 首先训练词向量, 然后将文本中的词向量相加, 作为文本的向量表示, 并应用于后续的文本分类中.

针对自然语言句子, 可以直接利用预训练的词向量, 也可以采用类似词向量的获取机制, 实现句子嵌入 (sentence embedding), 句子中的词序可以考虑也可不做考虑. 代表性的句子的向量化表示方法有 5 种: (1) 基于神经网络词袋模型. 句子做分词处理, 将包含在句子中的词对应的词向量进行某种加权计算, 最简单的是取所有词向量的平均值, 作为句子的向量表示, 方法简单但丢失了词序信息, 长文本比较有效, 短文本难以捕获语义组合信息; (2) 基于递归神经网络 RecNN (recursive neural network). 将句子按照某种拓扑结构, 比如句法树, 进行分解, 将结构中词对应的词向量递归处理得到句子表示, 方法易于理解, 但给定拓扑结构限制了使用范围; (3) 基于循环神经网络 RNN. 将句子看作时间序列, 表示成一个有顺序的向量序列, 通过对这个向量序列进行变换 (transformation) 和整合 (aggregation), 计算出对应的句子向量表示; (4) 基于卷积神经网络 CNN. 将句子看作符号序列, 通过多个卷积层和子采样层对文本序列进行处理, 得到一个固定长度的向量; (5) 综合或改进方法. 目前的做法是综合这些方法的优点, 结合具体任务, 进行模型选择或组合, 改进模型例如长短时记忆模型 LSTM、双向循环神经网络 Bi_RNN (bi-directional recurrent neural network) 等. 另外, 一些新的方法, 比如自回归建模、自编码建模、基于注意力机制建模、乱序语言模型建模等逐渐提出并获得了较好的效果[46,52].

对于段落或篇章表示, 可以借鉴句子向量的实现方法, 也可以采用层次化方法, 先获取句子表示, 然后以句子表示为输入得到段落或篇章表示 (paragraph/text embedding), 代表性方法有 3 种: (1) 基于卷积神经网络 CNN. 利

用卷积神经网络对句子建模, 以句子为单位再卷积和池化, 得到篇章表示; (2) 基于循环神经网络 RNN. 采用循环神经网络对句子建模, 然后再用循环神经网络建模以句子为单位的序列, 得到篇章表示; (3) 混合模型. 先用循环神经网络对句子建模, 然后以句子为单位再卷积和池化, 得到篇章表示. 循环神经网络及其各种变形相对适合处理文本序列, 应用较多.

近年来, 基于神经网络的深度学习研究和应用越来越广泛. 利用深度学习进行自然语言处理, 实现文本表示, 本质上是一种特征学习方法, 借助海量文本, 解决了复杂的“特征工程”问题, 反映了语义信息. 目前有多种神经网络模型用于建模文本, 比如利用卷积神经网络抽取部分单词作为输入特征, 类似于 n-grams 的思想, 利用循环神经网络模型具有时序特征的记忆性, 可按顺序将词向量特征输入[80−82]. 目前, 基于深度神经网络的文本表示代表性的工作包括:

(1) ELMo (embeddings from language models)

ELMo[83]实现了一词多义、动态更新的词嵌入建模. 先在一个大的语料库上训练语言模型, 得到词向量和神经网络结构, 接着进行领域转换 (domain transfer), 用训练数据来调优预训练好的 ELMo 模型, 这种训练数据的上下文信息就是词的语境. 在 ELMo 中, 利用双向 LSTM 语言模型, 目标函数取这两个方向语言模型的最大似然. ELMo 的本质思想是事先用语言模型学好一个单词的词向量, 此时多义词无法区分. 实际使用词向量的时候, 再根据上下文单词的语义去调整该单词的词向量表示, 经过调整后的词向量更能表达在这个具体上下文中的含义. ELMo 采用典型的两阶段过程: 第 1 个阶段是利用语言模型进行预训练; 第 2 个阶段是在做特定任务时, 从预训练网络中提取对应单词的词向量作为新特征补充到下游任务中. 基于 RNN 的 LSTM 模型训练时间长, 特征提取需要提升等是 ELMo 模型优化和提升的关键.

(2) Transformer/self-attention

Transformer[84]是 Google 提出的一种文本表示全新架构模型, 用来解决 LSTM 文本建模长距离依赖缺陷的问题. 模型采用 encoder-decoder 结构 (详细见第 2.5.4 节), encoder 将输入序列转化为一个上下文向量, 然后将其传递给 decoder, decoder 生成输出序列. 在 encoder 和 decoder 中不再采用 RNN、CNN 和 LSTM, 而是采用特征学习能力更强的“自注意力 (self-attention) 和多头注意力 (multi-head self-attention)”机制. 为充分发掘深度神经网络的特性, transformer 可以增加到非常深的深度, 以提升模型准确率. Transformer 是第一个采用纯 attention 搭建的模型, 在实现文本 seq2seq 过程中, encoder 和 decoder 过程都采用了堆叠的深度神经网络. Transformer 由于性能优良, 在自然语言处理中获得了广泛应用, 其存在的主要问题是训练中的序列上下文需要固定长度. 为解决这个限制, Transformer-XL[85]引入片段级递归机制 (segment-level recurrence mechanism) 和相对位置编码方案 (relative positional encoding scheme) 两种技术. 实验证明, Transformer-XL 几乎是目前最优 (SoTA) 的语言表示模型.

(3) Open AI GPT (generative pre-training)

GPT[86]的目标是为自然语言处理学习一个通用的文本表示, 采用了 12 层的深度神经网络, 能够在大量任务上进行应用. GPT 的预训练过程与 ELMo 类似, 主要的不同点有两个: (1) 特征抽取不是使用双层双向 LSTM, 而是使用 Transformer, 前面提到过 Transformer 的特征抽取能力要强于双层双向 LSTM; (2) GPT 的预训练虽然仍然是以语言模型作为目标任务, 但是采用的是单向的语言模型, 只采用目标词的上文 (context-before) 来进行预测. GPT-2 作为 GPT 的升级版本, 采用 48 层单向 Transformer (BERT 最深 24 层), 海量的超大数据集, 共有 15 亿个参数[87]. 最新的 GPT-3 模型在更庞大的数据集上进行训练, 包含的参数比 GPT-2 多了两个数量级, 达到 1 750 亿个, 比 GPT-2 有了极大的改进.

(4) BERT (bidirectional encoder representation from transformers)

BERT[88]是一种非常成功的文本表示学习模型, 即通过一个深层模型来学习文本特征, 这个模型可以从无标记数据集中预训练得到. 训练这个模型来预测句子中被 mask 的单词. 给定输入字符序列, 字符的某一部分用特殊符号 [mask] 替换, 训练模型从中恢复原始字符. BERT 直觉上要做更深的模型, 需要设置一个比语言模型更难的任务, 而实际上 BERT 选择了完形填空和句对预测这两个看似简单的任务来做预训练. 为解决完形填空任务中需要 mask 掉一些词带来的缺陷, BERT 通过随机的方式, 即 80% 加 mask, 10% 用其他词随机替换, 10% 保留原词的方

式, 这样模型就具备了迁移能力. BERT 模型是非图灵完备的, 无法单独完成 NLP 中的预测任务, 但一个好的表示会使得后续的任务更简单.

预训练语言表示分为基于特征的方法 (ELMo 为代表) 和基于微调 (Open AI GPT 为代表) 的方法. BERT 最重要的意义不在于模型选择和训练方法, 而是提出了一种全新的思路, 效果好且具备广泛的通用性, 绝大部分自然语言处理任务都可以采用类似的两阶段模式直接去提升效果.

(5) XLNet (extra long net)

GPT 和 BERT 的出现, 使自然语言处理任务的主流做法变为预训练 & 微调 (pre-train+finetune) 的形式, 先在大规模语料库上进行有监督或无监督预训练, 然后针对特定任务对模型微调. GPT 是一种自回归 (autoregressive, AR) 预训练模型, 而 BERT 是一种自编码器 (autoencoding, AE) 预训练模型, 两者的性能相当并均取得了良好的效果, 但都存在一定的问题. GPT 从前往后预测, 只能利用文本的单向信息, 无法做到挖掘上下文之间关系. BERT 解决了上下文依赖的问题, 存在的问题主要有两个: (1) 预训练任务和微调任务之间可能存在不一致; (2) 预测的时候, 多个 [mask] 之间的文本内容是相互独立的, 在预训练的时候对于依赖关系的挖掘不够充分.

XLNet[89]预训练过程采用双流自注意力机制, 模型中每层设两个 Transformer, 一个查询流 query stream 和一个内容流 context stream, 分别得到两组隐向量, 查询表示向量 g 和内容表示向量 h. 在预测下一个文本对象 (Unit) 的时候, 必须要包含该 Unit 的位置信息, g 向量包含了当前 Unit 所依赖的上下文信息, 以及该 Unit 的位置, 但是不包含该 Unit 自身的信息, 否则就会出现信息泄露. h 既包含 Unit 依赖的上下文信息, 也包含当前 Unit 的信息. 在模型最后一层, 对每个 g 向量, 预测该向量所表示的 Unit. XLNet 实质上是 ELMO、GPT 和 BERT 两阶段模型的延伸, 将自回归语言机制引入双向语言模型, 预训练模式符合下游任务序列结果的生成, 在生成类 NLP 任务中可以直接通过引入 XLNet 来改进效果.

归纳起来, 基于深度学习的文本表示的相关技术演进, 代表性的模型包括: NNLM (2003)、Word2Vec (2013); GloVe、Seq2Seq (2014); attention mechanism、memory-based neural networks (2015); fastText (2016); Transformer (2017); GPT-1、ELMO、BERT (2018); Transformer-XL、GPT-2、改进的 BERT 模型 (RoBERTa、ALBert、TinyBert、DistilBert、SpanBert、mBert、BART 等)、T5、XLNet、ERNIE (2019); GPT-3、ELECTRA (2020) 等, 这些模型为文本表示和自然语言处理提供了新的机遇. 基于神经网络的文本表示模型的主要优点包括[90,91]: (1) 数据驱动 (data driven), 与当前大数据时代相适应, 大量的语料易于获取; (2) 自动特征获取, 克服了繁琐的特征工程中的一些问题; (3) 神经网络在图像、视频等领域中的成功应用, 为 NLP 提供了良好的借鉴; (4) 不同粒度、不同领域的文本借助向量获得统一的文本表示形式, 简化了表示过程和后续任务的针对性处理; (5) 基于神经网络的系统易于计算机的并行处理, 实现了端到端的应用, 克服了传统 NLP 系统表示、任务两步法的弊端, 有较低的维护成本和较小的推理代价. 基于神经网络的文本表示研究存在的问题包括: (1) 神经网络模型可解释性差, 理论依据不足; (2) 深度学习需要海量训练数据, 不同数据集、不同数据规模、不同模型和不同的超参设置都会导致结果不同; (3) 文本表示建立在基于上下文建模的词向量基础上, 如何利用外部先验知识优化词向量表示模型, 以便提高文本特征表示能力; (4) 对词向量及其他粒度文本表示的质量评价. 基于神经网络的文本表示模型既有比较明显的优点, 也存在一些问题. 针对这些问题, 需要进一步研究、提高和改进, 主要包括: (1) 可解释性. 研究和探索神经网络模型应用于文本表示的理论基础和模型架构; (2) 组合性. 如何利用神经网络学习到一个好的不同粒度文本的嵌入表示, 如何基于已有嵌入表示建立更大粒度文本内容对象的表示模型[92]; (3) 适应性. 针对多领域、多语言、多模式等的具体 NLP 任务, 充分借鉴和融合相关技术.

2.5 基于表示学习的方法

人类在学习一个复杂概念时, 常规的思路是化繁为简, 这种思路反映在机器学习上, 如果原始数据经过提炼有更好的表达, 往往会使后续任务易于处理, 这实际上就是表示学习 (representation learning), 即找到对于原始数据更好的表达, 以便于后续任务的解决. Bengio 等人[40]将表示学习定义为: “学习数据的表征, 以便在构建分类器或其他预测器时更容易提取有用的信息”. 表示学习也称学习表示, 试图通过机器学习算法自动地进行特征空间映射, 获得原始数据的有效表示, 最终提高机器学习模型的性能. 表示学习涉及到两个核心问题: (1) 什么是好的表示?

(2) 怎么学习到好的表示?

对自然语言处理来说, 源文本数据的底层特征和高层认知语义信息之间的差异性和不一致性造成了语义鸿沟问题. 自然语言处理长期以来的一个重要挑战就是从无结构的文本中提取特征, 不论是基于规则的系统还是基于统计的技术, 都需要获取文本的特征, 将输入信息转换为有效的特征[35]. 传统基于人工处理的特征工程成本高、效率低, 针对自然语言的表示学习近年来取得了很多进展, 表示学习使得不同语种、不同任务、不同领域的迁移学习 (transfer learning) 和领域自适应 (domain adaption) 任务取得了成功[93,94]. 自然语言处理中好的文本表示是一个主观性的概念, 没有明确的标准, 一般应具备以下几点: (1) 表示能力. 表示模型具有一定的深度, 同样规模的表示度量, 比如向量、矩阵等, 可以包含更多的文本信息. (2) 易用性. 文本的表示容易实现, 使后续文本处理任务简单, 性能得到提升. (3) 泛化能力. 好的文本表示应该具有一般性, 是任务或领域独立的, 可以较容易地迁移到其他任务上. 判定一种表示优于另一种, 需要具体问题具体分析, 对于特定任务来说, 通常依赖于后续任务的效率和性能比较[95].

前面提到, 文本表示学习一定程度上解决了不同粒度的文本、不同任务的自然语言处理以及跨领域文本的特征获取问题, 将文本潜在语法或语义特征分布式地存储在稠密、连续、低维的向量中. 文本表示学习的实现基于神经网络, 与深度学习、元学习、多任务学习、迁移学习、图论、注意力机制等关系密切. 不同粒度的语言表示有不同的用途, 如字、词和短语表示主要用于预训练, 服务于下游任务, 句子、段落、篇章表示可直接用于文本分类、阅读理解等具体任务.

2.5.1 文本表示中的深度学习和表示学习

前面提到的深度学习, 其基本单元是向量, 将文本建模对象对应到一组向量 $X(x^{(1)},x^{(2)},\cdots,x^{(n)})$, 然后通过多层的神经网络变换、整合得到若干新向量 $H(h^{(1)},h^{(2)},\cdots,h^{(m)})$, 每个向量 $h^{(i)}$ 就是对文本建模对象的特征表示, 是一个借助算法逐层抽象学习的训练过程, 最后基于 H 得到输出 y. 表示学习将许多深度学习形式联系在了一起, 关键环节是构建具有一定深度的多层次特征表示. 深度学习是表示学习最好的体现, 算法自动构建特征, 解决了人工特征构建的问题[96,97]. 所以有观点认为, 深度学习就是表示学习, 至少是表示学习的经典代表. 表示学习与深度学习的关系如图 7 所示.

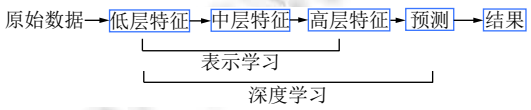

图 7 表示学习和深度学习关系示意图

表示学习在自然语言处理中的应用, 主要包括对文本的无监督/有监督预训练、分布式表示以及迁移学习等. 其中, 无监督学习的代表是贪心逐层无监督预训练模型 (greedy layer-wise unsupervised pretraining), 包括受限玻尔兹曼机 (restricted Boltzmann machine, RBM)、单层自编码器、稀疏自编码器、堆叠自编码器 (stacked auto-encoder, SAE) 等[98,99]. 分布式表示的常用例子是词嵌入. 有监督学习一般通过丢弃 (Dropout) 策略或批标准化来正则化, 在极大的标注数据集上, 比无监督预训练效果要好, 很多任务的性能能够达到人类级别. 这里所说的预训练, 实际上就是指迁移学习.

2.5.2 文本表示中的图嵌入和表示学习

21 世纪初提出的图嵌入 (graph embedding) 方法, 也称网络嵌入 (network embedding), 通过保留图中节点信息和节点之间边的拓扑结构, 将图中节点表示为低维向量空间, 以便后续的机器学习算法进行处理, 这种方法进一步发展, 统称为图表示学习 (graph representation learning, GRL). 图神经网络 (graph neural network, GNN)[100]是与之相近的概念, 也可认为是图表示学习的高阶版本, 实质上都是深度学习应用于图表示学习的具体实现, 一般不做严格区分. GNN 是一种建立在图模型基础上的基于神经网络表示学习的连接模型, 以不动点理论为基础, 一般定义为: 假设网络 $G=(V,E)$, 其中, $V=\{v_i\}$ 是节点集合, $E=\{(v_i,v_j)\}$ 是边的集合, GNN 网络表示将学习到网络中每个节点在低维空间的向量表示 $\Phi(v_i)$. 其中, $\Phi(v_i)\in\mathbb{R}^k$, $\Phi\in\mathbb{R}^{|V|\times k}$, 且 $k<<|V|$, $\mathbb{R}$ 为实数空间, k 为低维特征空间的维度.

从定义可以看出, GNN 是将图中每个节点都映射到一个低维向量空间, 将复杂、高维的图数据转化成低维稠密的向量, 并且在空间内保持原图结构中的节点信息和结构信息[101]. 2010s 年代以来, GNN 研究在基本的图向量

化实现的基础上, 重点关注网络稀疏性解决策略和可伸缩图嵌入的技术实现. 图表示学习方法分为基于因子分解的方法、基于随机游走的方法和基于深度学习的方法. 基于因子分解的文本图表示学习主要针对文本图表示的邻接矩阵进行近似分解, 并将分解的结果作为文本特征的嵌入表示, 包括 LLE (local linear embedding)、Laplacian Eigenmaps、Cauchy graph embedding 等模型. 基于随机游走的文本图表示学习中通过节点之间的消息传递来捕捉文本的依赖关系, 基本思想是基于节点的局部邻居信息对节点进行嵌入表示学习, 简言之就是通过神经网络来聚合每个节点及其周围节点的信息. 这种方法实质上是一种线性化思维, 通过遍历生成线性序列, 然后借鉴 Word2Vec 的实现机制在线性序列上进行处理. 基于深度学习的文本图表示学习是将深度学习模型迁移到图数据上, 进行端到端的建模, 包括图卷积神经网络[102]、基于注意力机制的 GNN、生成式 GNN 等. GNN 方法分为监督和无监督两种方法, 代表性的无监督的方法包括 node2vec、DeepWalk 等随机游走方法[73], 训练模型使得相似的节点具有相似的嵌入.

图模型和神经网络作为文本表示的主流建模方法, 都可以看作是一种网络结构, 两者有许多相似之处, 结合也越来越密切. 一方面可以利用神经网络的抽象表示能力建模并实现图模型中的推断问题, 比如变分自编码器、生成对抗网络或势能函数等; 另一方面可以利用图模型算法来解决神经网络中的学习和推断问题, 比如图神经网络 GNN[46]. 建立在深度学习基础上的文本图表示学习, 实际上是表示学习的一种, 但两者也有很大的不同. 图模型和神经网络模型的简单比较如表 2 所示.

表 2 图模型和神经网络模型的比较

项目	图模型	神经网络模型
节点	随机变量, 明确的解释	神经元/计算节点, 无明确解释
边(联系)	变量之间依赖关系, 稀疏, 人工定义	无显示的依赖关系, 可以堆叠
模型作用(好处)	统计推断	非线性变换
模型类别	判别/生成模型	判别模型
目标函数(模型参数学习)	似然函数或条件似然函数、EM算法(包含隐变量)	损失函数, 比如交叉熵或平方误差等

近年来提出的生成对抗网络 (generative adversarial networks, GAN) 在文本表示学习中也引起了极大的关注[103,104]. GAN 包含两部分: 生成器用来生成尽可能真实的自然语言文本, 去“欺骗”或“误导”判别器; 判别器尽最大努力甄别真实语言文本与生成的文本. 训练 GAN 就是使生成器和判别器相互博弈, 达到真实文本和生成器生成的文本难以区分的效果. 比较代表性的改进模型有 SegGAN、GraphGAN、ANE 等[105−108].

采用生成模型的基于图的文本表示学习, 是对每个节点、每条边, 在文本中找到一个潜在的、真实的连续性分布, 也就是通过学习获得节点和边的 embedding. 采用判别模型的基于图的文本表示学习, 一般将两个节点作为联合的特征, 预测两者之间边的概率. 也可以将生成模型和判别模型结合起来进行学习[109,110].

2.5.3 文本表示中的注意力机制和表示学习

受人类视觉机制的启发, 注意力机制首先在图像领域取得了成功, 最近几年注意力模型在深度学习的各个领域被广泛使用. 目前, 自然语言处理任务的表示学习几乎都要用到了注意力模型, 大多数注意力机制都是在深度学习的常见编码-解码 (encoder-decoder) 框架上发挥作用的, encoder-decoder 框架适合处理由一个文本序列到另外一个文本序列的处理模型, 也叫做序列-序列学习 (sequence to sequence learning)[111]. 常用的 encoder-decoder 框架的抽象表示如图 8 所示.

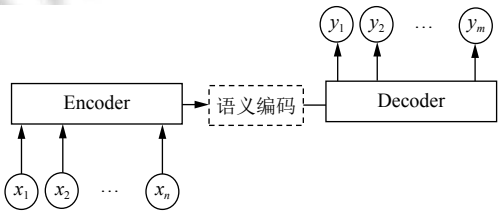

图 8 编码-解码框架

编码-解码框架中的编码神经网络一般是多层的 CNN、RNN、LSTM 等, 将输入序列编码成一个固定长度的向量 C, 类似的, 解码框架也是神经网络框架, 将之前生成的固定向量 C 再解码成输出序列. 以机器翻译为例,

$X(x_1,x_2,\cdots,x_n)$ 是汉语句子, $Y(y_1,y_2,\cdots,y_m)$ 是对应的英文翻译, 给定输入 X, 通过编码-解码框架进行翻译, X 和 Y 分别由各自语言的单词序列组成. encoder-decoder 框架简单、易于理解, 但是局限性也大, 语义向量 C 可能会失去整个序列的部分信息, 特别是对长输入序列, 由此解码的准确度就会下降.

注意力机制最核心的工作就是在序列的不同时刻产生不同的语言编码向量, 量化要重点关注输入序列中的哪些部分, 然后根据关注区域产生后续的输出. 形象化的模型表示如图 9 所示.

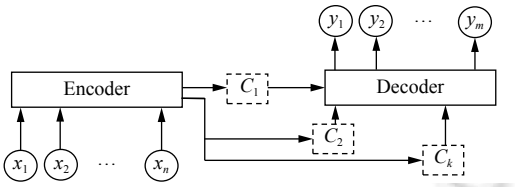

图 9 基于注意力机制的编码-解码框架

引入注意力机制的 encoder-decoder 模型, 编码器将输入信息编码成一个向量的序列, 序列中不同片段的重要性反映在不同的向量中. 解码的时候, 每一步选择向量序列中的一个子集进行下一步处理, 这样的每个输出, 都能够充分利用输入序列携带的信息[112−114].

注意力机制具有直观性、通用性和可解释性, 可以进行不同应用领域的特征表示和语言建模, 近年来成为一个活跃的研究领域. 另外, 一些有趣的方向, 包括更平滑地整合外部知识库、训练前嵌入和多任务学习、无监督学习、稀疏性学习和原型学习等, 本文不再展开讨论.

3 其他文本表示方法分析

3.1 基于事件的方法

事件是现实世界中经常提及的一个词汇, 涉及的动作、对象、时间、环境等信息伴随事件的发生而存在, 伴随事件的结束而撤销. 事件是客观的、是不依赖特定语言的对现实的抽象. 从自然语言处理角度, 文本中描述事件及其信息的符号合理有效地表示事件、事件与事件之间的关系, 以文本中的事件为单位进行文本表示尽管并未列在文本表示的宏观分类列表上, 但仍具有一定的可行性和实用性, 特别是针对特定文体的自然语言处理, 比如叙事文本、新闻文本等. 从认知科学角度, 事件是人类认识和理解现实世界、描述或传播信息的基本单元, 以事件为知识表示单元对文本中的事件以及事件关系进行有效表示是一项基础性工作, 可以为事件本体以及基于事件的知识推理提供服务[115,116].

3.2 基于语义的方法

自然语言处理的核心思想是对语义的表示和分析, 对文本语义的表示简单来说就是将无结构文本表示成结构化的形式. 首先形成语义对象, 然后分析语义对象之间的联系或整个表达的语义结构, 这和常规意义上的文本表示是相似的, 上文提到的词向量、句向量、主题模型等都可认为是语义表示的形式[117,118]. 除此之外, 还包括逻辑、框架理论、语言组成和语境理论等. 事实上, 语义分析是指运用各种方法, 学习与理解一段文本所表示的语义内容, 任何对语言的理解都可以归结为语义分析的范畴. 根据文本组成粒度的不同, 语义分析又可进一步分解为词汇级、句子级以及篇章级语义分析. 词汇级语义分析关注词的语义, 句子级语义分析关注整个句子所表达的语义, 篇章语义分析旨在研究自然语言文本的内在结构并理解文本不同粒度间的语义关系. 简言之, 语义分析的目标就是通过建立有效的模型和系统, 实现各个语言单位 (包括词汇、句子和篇章等) 的自动语义分析, 从而实现对整个文本的真实语义的理解. 代表性的基于语义的文本表示方法包括[119,120]: (1) 基于语境框架的文本表示, 以概念层次网络理论为基础, 将文本信息抽象成领域信息、情景信息、背景信息 3 方面, 利用一个包含语境三要素的三维信息空间描述来进行文本表示. (2) 基于本体论的文本表示, 通过领域本体将文本的特征词映射为概念, 通过对文本概

念进行处理以完成文本表示过程.

3.3 基于知识的方法

知识表示、知识应用一直以来都是人工智能研究的重要问题, 一方面, 借助自然语言处理技术可以从自然语言文本中抽取出知识, 另一方面, 知识, 特别是领域知识对自然语言处理具有辅助性作用. 尽管现在主流自然语言处理系统中的文本表示较少关注知识, 甚至都不关注语言学知识, 但针对特定应用的领域知识和语言学知识对自然语言处理有极大的帮助作用, 语言的领域应用不仅需要静态知识, 更需要动态知识. 领域专家的思维、决策知识等加以整合和表示, 赋予机器, 从而一定程度上降低对专家的依赖[121−123]. 举一个例子, 前面提到的基于上下文训练词向量依赖于大量静态文本, 结合结构化知识的想法就是利用类似 WordNet、HowNet、Wikipedia 等作为监督, 辅助词向量的训练, 同义词信息、词语间的关系 (同义, 上下位等) 表示成函数约束以提升词向量的准确度[124].

4 文本表示发展趋势

文本表示作为自然语言处理的基础, 伴随着语言学、认知科学和人工智能的发展而进步, 其方法从早期的规则法, 到基于统计的方法, 再到近年来的深度学习方法和表示学习, 从早期的离散表示, 到基于矩阵分解的分布式表示, 再到基于神经网络的分布式表示, 从特定任务中字、词、句子、篇章的表示, 再到端到端大规模自动文本表示学习, 给自然语言处理任务带来了越来越大的便利. 考虑文本表示面临的挑战和需要进一步解决的问题, 包括: 语言中出现的所有符号是否都需要表示, 特别是无意义符号; 新词以及低频词的表示学习方法; 篇章中复杂语义的有效表示; 目前的基于向量的数据结构之外是否有更好的表示结构等. 随着深度学习、神经科学和脑科学、数据挖掘等技术的发展, 为了更好地完成自然语言处理任务, 文本表示将会进一步研究并得到更大程度的解决, 其发展趋势归纳为以下 7 方面.

(1) 文本表示与知识的融合. 知识应用于文本表示既是可理解的, 又能缓解计算资源限制, 特别是对于垂直领域, 专业知识非常有效, 如医疗、金融等文本表示. 构建语言表示和知识的联系, 如何利用已有的知识库来改进词嵌入模型, 结合知识图谱和未标注语料学习知识和词向量表示, 更有效地实现文本嵌入.

(2) 跨语种的语言统一表示研究. 借助脑科学、神经科学和认知科学的发展成果, 进一步认识世界以及人类思想的表达能力, 探索不同语种的语言符号及语法规则的特点, 分析其语言和文本表示的相似性, 考虑将相同语义的不同语言的文本进行相同或相近的表示, 试图为不同语种构建统一的语言表示模型, 提高各语种语言的表示能力.

(3) 多粒度文本的联合表示. 以语言研究为基础, 分析语言本身的层次结构, 分析不同粒度文本之间的关系, 构建多粒度文本的联合语义表示模型.

(4) 少资源文本表示学习. 目前, 文本表示不论是统计方式还是基于神经网络, 一般需要大量语料的学习, 受限于数据和算法, 大多数文本表示方法经常过滤掉少资源或低频的词, 即使处理这些词, 也难以很好地建模, 这样往往会丢失有价值的信息, 降低表示能力. 人类对少资源或低频词的学习常常通过字典或少量语言样本进行, 因此, 研究如何通过少量观察样本来学习新词和低频词的表示, 是文本表示研究的方向之一.

(5) 文本表示的自动学习. 近年来, 由于深度学习对非结构数据的强大表示和学习能力, 表示学习在图像、视频、语音等领域取得了不错的效果. 实用化的自然语言处理需要对复杂语境进行建模, 基于海量文本数据自动学习文本的表示, 必然是自然语言表示研究的一个重要方向.

(6) 多模态深度语义的融合表示. 为实现对信息的多维度和深层次理解, 声音、图像、视频、文本等不同模态的信息综合感知和认知表示建模, 拟合人类学习过程[125−129]. 在安防、教育、医疗、智慧城市、金融等行业进行广泛应用.

(7) 传统表示模型的深入研究和广泛应用. 伴随着自然语言处理发展过程而提出的经典文本表示模型, 包括基于文法的模型、基于统计的模型和基于图的模型等, 这些方法理论基础成熟、模型简单有效, 已经广泛应用于许多自然语言处理任务中. 对这些方法进行深入研究、优化和融合, 一方面能快速实现具体的自然语言处理任务, 另一方面也能降低模型对时间、资源等的过度依赖[130−132].

5 总结和进一步讨论

从自然语言处理技术层面上, 再一次探讨文本表示与机器学习, 特别是深度学习的关系. 统计自然语言处理系统通常由训练数据和统计模型两部分组成, 传统机器学习方法的数据获取存在标注代价高、规范性差和数据稀疏等问题, 模型需要的特征存在获取困难的问题[133]. 深度学习尽管目前存在理论基础不足、可解释性较差等问题, 但其建立在多层非线性变换的神经网络结构之上, 核心目标是对特征的抽象和学习. 深度学习把自然语言处理的问题域从离散符号域转换到了连续数值域, 使问题定义和支撑工具发生了改变, 极大地促进了自然语言处理的发展. 深度学习为自然语言处理带来的变化包括[134,135]: (1) 使用统一的分布式向量表示不同粒度的语言对象, 如词、短语、句子和篇章等; (2) 使用卷积、循环、递归等神经网络模型对不同语言对象的向量表示进行组合, 获得更大语言对象的表示; (3) 使用统一的模型处理和融合多模态信息, 不同语种的自然语言以及图像、语音、视频等不同模态的数据都可以通过类似的方式在相同的向量空间中表示, 然后通过向量运算实现推理、生成等任务并应用于其他相关任务中. 目前和今后较长一段时间, 以 Word2Vec[36]、GloVe[77]、fastText[79]为代表的嵌入式表示方法, 以 Transformer[84]、GPT[86,87]、BERT[88]、XLNet[89]为代表的自然语言深度学习框架, 以预训练加微调为代表的自然语言处理基本流程将成为进一步研究和发展的主导方向[136−138]. 主要的技术趋势就是利用更大容量的模型, 近于无限的无监督文本, 编码语言学知识和人类知识. 另外, 自然语言处理与大数据工程、知识库、脑科学和认知科学等的融合发展和相互启发研究, 将推动包括自然语言处理在内的人工智能技术迈向新的台阶.

References:

[1] Ray J, Johnny O, Trovati M, Sotiriadis S, Bessis N. The rise of big data science: A survey of techniques, methods and approaches in the field of natural language processing and network theory. Big Data and Cognitive Computing, 2018, 2(3): 22. [doi: 10.3390/bdcc2030022]

[2] Kim Y, Lee J, Lee EB, Lee JH. Application of natural language processing (NLP) and text-mining of big-data to engineering-procurement-construction (EPC) bid and contract documents. In: Proc. of the 6th Conf. on Data Science and Machine Learning Applications (CDMA). Riyadh: IEEE, 2020. 123–128. [doi: 10.1109/CDMA47397.2020.00027]

[3] Friederici AD, Chomsky N, Berwick RC, Moro A, Bolhuis JJ. Language, mind and brain. Nature Human Behaviour, 2017, 1(10): 713–722. [doi: 10.1038/s41562-017-0184-4]

[4] Yu J. Language and Turing test. Acta Automatica Sinica, 2016, 42(5): 668–669 (in Chinese with English abstract). [doi: 10.16383/j.aas.2016.y000004]

[5] Desaussure F. Course in General Linguistics. New York: McGraw-Hill, Inc., 1965.

[6] Liang JY, Liu HT. Interdisciplinary studies of linguistics: Language universals, human cognition and big-data analysis. Journal of Zhejiang University (Humanities and Sciences), 2016, 46(1): 108–118 (in Chinese with English abstract). [doi: 10.3785/j.issn.1008-942X.CN33-6000/C.2015.10.231]

[7] Lai YY, Li C, Goldwasser D, Neville J. Better together: Combining language and social interactions into a shared representation. In: Proc. of the TextGraphs-10: The Workshop on Graph-based Methods for Natural Language Processing. San Diego: Association for Computational Linguistics, 2016. 29–33. [doi: 10.18653/v1/W16-1405]

[8] Sun MS, Liu T, Ji DH, Sui ZF, Zhao J, Zhang B, Wushouer S, Yu SW, Zhu J, Li JM, Liu Y, Wang HF, Turgun I, Liu Q, Liu ZY. Frontiers of language computing. Journal of Chinese Information Processing, 2014, 28(1): 1–8 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2014.01.001]

[9] Zong CQ. Statistical Natural Language Processing. 2nd ed., Beijing: Tsinghua University Press, 2013. 8 (in Chinese).

[10] Taskin Z, Al U. Natural language processing applications in library and information science. Online Information Review, 2019, 43(4): 676–690. [doi: 10.1108/OIR-07-2018-0217]

[11] Chu XM, Zhu QM, Zhou GD. Discourse primary-secondary relationships in natural language processing. Chinese Journal of Computers, 2017, 40(4): 842–860 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2017.00842]

[12] Li SC. General trend of the change in world linguistic academic thinking. Dongyue Tribune, 2017, 38(8): 163–168 (in Chinese with English abstract). [doi: 10.15981/j.cnki.dongyueluncong.2017.08.022]

[13] Zhao YG. Explanations on the evolution of language and the forward road of bio-linguistic: Concurrent comments on why only us: Language and evolution. Academic Exploration, 2018, (6): 107–116 (in Chinese with English abstract). [doi: 10.3969/j.issn.1006-723X.2018.06.016]

[14] Feng ZW. Formal Models of Natural Language Processing. Hefei: China University of Science and Technology Press, 2010. 1 (in Chinese).
[15] Zhang L, Wei NX. Local grammar: Evolution, status quo, and prospects. Contemporary Linguistics, 2018, 20(1): 103–116 (in Chinese with English abstract).
[16] Minsky M. Semantic Information Processing. Cambridge: MIT Press, 1968. 440–441.
[17] Schank RC. Conceptual Information Processing. Amsterdam: Elsevier Science Inc., 1975. 5–21.
[18] Berwick RC, Chomsky N. Why Only Us: Language and Evolution. Cambridge: MIT Press, 2015.
[19] Bendersky M, Croft WB. Modeling higher-order term dependencies in information retrieval using query hypergraphs. In: Proc. of the 35th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Virtual Event: Association for Computing Machinery, 2012. 941–950. [doi: 10.1145/2348283.2348408]
[20] Sidorov G. Syntactic N-Grams in Computational Linguistics. Cham: Springer, 2019. 3–86. [doi: 10.1007/978-3-030-14771-6]
[21] Barkovich A. Informational linguistics: Computer, internet, artificial intelligence and language. In: Proc. of the ICAIIC. 2019. 8–13.
[22] Liu J, Lin L, Ren HL, Gu MH, Wang J, Youn G, Kim JU. Building neural network language model with POS-based negative sampling and stochastic conjugate gradient descent. Soft Computing, 2018, 22(20): 6705–6717. [doi: 10.1007/s00500-018-3181-2]
[23] Bengio Y, Ducharme R, Vincent P, Janvin C. A neural probabilistic language model. The Journal of Machine Learning Research, 2003, 3: 1137–1155.
[24] Mikolov T, Karafiát M, Burget L, Černocký JH, Khudanpur S. Recurrent neural network based language model. In: Proc. of the INTERSPEECH. Chiba: ISCA, 2010. 1045–1048.
[25] Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling. In: Proc. of the INTERSPEECH. 2012. 194–197.
[26] Wang P, Xu JM, Xu B, Liu CL, Zhang H, Wang FY, Hao HW. Semantic clustering and convolutional neural network for short text categorization. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int'l Joint Conf. on Natural Language Processing (Volume 2: Short Papers). Beijing: Association for Computational Linguistics, 2015. 352–357. [doi: 10.3115/v1/P15-2058]
[27] Jing K, Xu JG, He B. A survey on neural network language models. arXiv: 1906.03591, 2019.
[28] Samuel S, Roehr-Brackin K, Pak H, Kim H. Cultural effects rather than a bilingual advantage in cognition: A review and an empirical study. Cognitive Science, 2018, 42(7): 2313–2341. [doi: 10.1111/cogs.12672]
[29] Siew CSQ, Wulff DU, Beckage NM, Kenett YN. Cognitive network science: A review of research on cognition through the Lens of network representations, processes, and dynamics. Complexity, 2019, 2019: 2108423. [doi: 10.1155/2019/2108423]
[30] Xu YJ. Wittgenstein, phenomenology and cognitive linguistics. Fudan Journal of the Humanities and Social Sciences, 2018, 11(2): 219–236. [doi: 10.1007/s40647-017-0182-y]
[31] Brenda M. A cognitive perspective on the semantics of near. Review of Cognitive Linguistics, 2017, 15(1): 121–153. [doi: 10.1075/rcl.15.1.06bre]
[32] Feiman R, Snedeker J. The logic in language: How all quantifiers are alike, but each quantifier is different. Cognitive Psychology, 2016, 87: 29–52. [doi: 10.1016/j.cogpsych.2016.04.002]
[33] Ali I, Melton A. Graph-based semantic learning, representation and growth from text: A systematic review. In: Proc. of the 13th Int'l Conf. on Semantic Computing (ICSC). Newport Beach: IEEE, 2019. 118–123. [doi: 10.1109/ICOSC.2019.8665592]
[34] Liu L, Chen J, Fieguth P, Zhao GY, Chellappa R, Pietikäinen M. From BOW to CNN: Two decades of texture representation for texture classification. International Journal of Computer Vision, 2019, 127(1): 74–109. [doi: 10.1007/s11263-018-1125-z]
[35] Liu ZY, Lin YK, Sun MS. Representation Learning for Natural Language Processing. Singapore: Springer, 2020. [doi: 10.1007/978-981-15-5573-2]
[36] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proc. of the 26th Int'l Conf. on Neural Information Processing Systems. Sydney: Curran Associates Inc., 2013. 3111–3119.
[37] Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Communications of the ACM, 1975, 18(11): 613–620. [doi: 10.1145/361219.361220]
[38] Xu TQ. Zi as the basic structural unit and linguistic studies. Language Teaching and Linguistic Studies, 2005, (6): 1–11 (in Chinese with English abstract).
[39] Zhang RN. Review of the research on teaching Chinese as a foreign language under the character-based theory. Modern Chinese, 2018, (4): 134–139 (in Chinese with English abstract).
[40] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798–1828. [doi: 10.1109/TPAMI.2013.50]
[41] Granados A. Analysis and study on text representation to improve the accuracy of the normalized compression distance. AI

Communications, 2012, 25(4): 381–384. [doi: 10.3233/AIC-2012-0529]

[42] Dourado ÍC, Galante R, Gonçalves MA, da Silva Torres R. Bag of textual graphs (BoTG): A general graph-based text representation model. Journal of the Association for Information Science and Technology, 2019, 70(8): 817–829. [doi: 10.1002/asi.24167]

[43] Harris ZS. Distributional structure. WORD, 1954, 10(2–3): 146–162. [doi: 10.1080/00437956.1954.11659520]

[44] Firth J. A synopsis of linguistic theory, 1930-55. In: Studies in Linguistic Analysis. Special Volume of the Philological Society. Oxford: Blackwell, 1957. 1–31.

[45] Wang SP, Cai JY, Lin QH, Guo WZ. An overview of unsupervised deep feature representation for text categorization. IEEE Trans. on Computational Social Systems, 2019, 6(3): 504–517. [doi: 10.1109/TCSS.2019.2910599]

[46] Qiu XP. Neural Networks and Deep Learning. Beijing: China Machine Press, 2020. 4 (in Chinese).

[47] Chew PA, Bader BW, Helmreich S, Abdelali A, Verzi SJ. An information-theoretic, vector-space-model approach to cross-language information retrieval. Natural Language Engineering, 2011, 17(1): 37–70. [doi: 10.1017/S1351324910000185]

[48] Tsatsaronis G, Panagiotopoulou V. A generalized vector space model for text retrieval based on semantic relatedness. In: Proc. of the 12th Conf. of the European Chapter of the Association for Computational Linguistics: Student Research Workshop. Athens: Association for Computational Linguistics, 2009. 70–78.

[49] Dong RF, Liu CA, Yang GT. TF-IDF based loop closure detection algorithm for SLAM. Journal of Southeast University (Natural Science Edition), 2019, 49(2): 251–258 (in Chinese with English abstract). [doi: 10.3969/j.issn.1001-0505.2019.02.008]

[50] Niu FG. Basic co-occurrence latent semantic vector space mode. Journal of Classification, 2019, 36(2): 277–294. [doi: 10.1007/s00357-018-9283-9]

[51] Hajjem M, Latiri C. Combining IR and LDA topic modeling for filtering microblogs. Procedia Computer Science, 2017, 112: 761–770. [doi: 10.1016/j.procs.2017.08.166]

[52] Liu ZY, Huang WY, Zheng YB, Sun MS. Automatic keyphrase extraction via topic decomposition. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP'10). Cambridge: Association for Computational Linguistics, 2010. 366–376.

[53] Ding ZY, Qiu XP, Qi Z, Huang XJ. Topical translation model for microblog hashtag suggestion. In: Proc. of the 23rd Int'l Joint Conf. on Artificial Intelligence. Beijing: AAAI Press, 2013. 2078–2084.

[54] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. The Journal of Machine Learning Research, 2003, 3: 993–1022.

[55] Pu XJ, Jin R, Wu GS, Han DY, Xue GR. Topic modeling in semantic space with keywords. In: Proc. of the 24th ACM Int'l on Conf. on Information and Knowledge Management. Melbourne: Association for Computing Machinery, 2015. 1141–1150. [doi: 10.1145/2806416.2806584]

[56] Siu MH, Gish H, Chan A, Belfield W, Lowe S. Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery. Computer Speech & Language, 2014, 28(1): 210–223. [doi: 10.1016/j.csl.2013.05.002]

[57] Schenker A, Last M, Bunke H, Kandel A. Graph representations for Web document clustering. In: Proc. of the 1st Iberian Conf. on Pattern Recognition and Image Analysis. Puerto de Andratx: Springer, 2003. 935–942. [doi: 10.1007/978-3-540-44871-6_108]

[58] Sonawane SS, Kulkarni PA. Graph based representation and analysis of text document: A survey of techniques. Int'l Journal of Computer Applications, 2014, 96(19): 1–8. [doi: 10.5120/16899-6972]

[59] Chen YY, Lu H, Qiu J, Wang L. A tutorial of graph representation. In: Proc. of the 5th Int'l Conf. on Artificial Intelligence and Security. New York: Springer, 2019. 368–378. [doi: 10.1007/978-3-030-24274-9_33]

[60] Zhao JS, Li Z, Zhu QM, Zhou GD. Extracting and analyzing social networks from Chinese literary. Journal of Chinese Information Processing, 2017, 31(2): 99–106, 116 (in Chinese with English abstract).

[61] Mihalcea R, Tarau P. TextRank: Bringing order into texts. In: Proc. of the EMNLP. Barcelona: ACM, 2004. 404–411.

[62] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: Bringing order to the Web. Stanford: Stanford InfoLab, 1999.

[63] Zhao JS, Zhu QM, Zhou GD, Zhang L. Review of research in automatic keyword extraction. Ruan Jian Xue Bao/Journal of Software, 2017, 28(9): 2431–2449 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5301.htm [doi: 10.13328/j.cnki.jos.005301]

[64] Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999, 46(5): 604–632. [doi: 10.1145/324133.324140]

[65] Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature, 1998, 393(6684): 440–442. [doi: 10.1038/30918]

[66] Barabási AL, Albert R. Emergence of scaling in random networks. Science, 1999, 286(5439): 509–512. [doi: 10.1126/science.286.5439.509]

[67] Cancho RFI, Solé RV. The small world of human language. Proc. of the Royal Society B: Biological Sciences, 2001, 268(1482):

2261–2265. [doi: 10.1098/rspb.2001.1800]

[68] Holovatch Y, Palchykov V. Complex networks of words in fables. In: Kenna R, MacCarron M, MacCarron P. Maths Meets Myths: Quantitative Approaches to Ancient Narratives. Cham: Springer, 2017. 159–175. [doi: 10.1007/978-3-319-39445-9_9]

[69] Balinsky H, Balinsky A, Simske SJ. Automatic text summarization and small-world networks. In: Proc. of the 11th ACM Symp. on Document Engineering. Limerick: ACM, 2011. 175–184. [doi: 10.1145/2034691.2034731]

[70] Pardo TAS, Antiqueira L, Nunes MDGV, Oliveira ON, Da Fontoura Costa L. Using complex networks for language processing: The case of summary evaluation. In: Proc. of the 2006 Int'l Conf. on Communications, Circuits and Systems. Guilin: IEEE, 2006. 2678–2682. [doi: 10.1109/ICCCAS.2006.285222]

[71] Lozano S, Calzada-Infante L, Adenso-Díaz B, García S. Complex network analysis of keywords co-occurrence in the recent efficiency analysis literature. Scientometrics, 2019, 120(2): 609–629. [doi: 10.1007/s11192-019-03132-w]

[72] Yan JH, Wang CY, Cheng WL, Gao M, Zhou AY. A retrospective of knowledge graphs. Frontiers of Computer Science, 2018, 12(1): 55–74. [doi: 10.1007/s11704-016-5228-9]

[73] Wang Q, Mao ZD, Wang B, Guo L. Knowledge graph embedding: A survey of approaches and applications. IEEE Trans. on Knowledge and Data Engineering, 2017, 29(12): 2724–2743. [doi: 10.1109/TKDE.2017.2754499]

[74] Liu H, Zhang YZ, Wang YP, Lin Z, Chen YG. Joint character-level word embedding and adversarial stability training to defend adversarial text. Proc. of the AAAI Conf. on Artificial Intelligence, 2020, 34(5): 8384–8391. [doi: 10.1609/aaai.v34i05.6356]

[75] Hinton GE. Learning distributed representations of concepts. In: Proc. of the 8th Annual Conf. of the Cognitive Science Society. Hillsdale: Lawrence Erlbaum, 1986. 1–12.

[76] Levy O, Goldberg Y. Neural word embedding as implicit matrix factorization. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2177–2185.

[77] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Doha: Association for Computational Linguistics, 2014. 1532–1543. [doi: 10.3115/v1/D14-1162]

[78] Le Q, Mikolov T. Distributed representations of sentences and documents. In: Proc. of the 31st Int'l Conf. on Machine Learning. Beijing, 2014. II-1188–II-1196.

[79] Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. arXiv: 1607.01759, 2016.

[80] Vinyals O, Le Q. A neural conversational model. arXiv: 1506.05869, 2015.

[81] Shen DH, Wang GY, Wang WL, Min MR, Su QL, Zhang YZ, Li CY, Henao R, Carin L. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne: Association for Computational Linguistics, 2018. 440–450. [doi: 10.18653/v1/P18-1041]

[82] Peters ME, Ammar W, Bhagavatula C, Power R. Semi-supervised sequence tagging with bidirectional language models. arXiv: 1705.00108, 2017.

[83] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. arXiv: 1802.05365, 2018.

[84] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conferenc. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[85] Dai ZH, Yang ZL, Yang YM, Carbonell J, Le QV, Salakhutdinov R. Transformer-XL: Attentive language models beyond a fixed-length context. arXiv: 1901.02860, 2019.

[86] Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. https://www.cs.ubc ca/~amuham01/LING530/papers/radford2018improving.pdf.

[87] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

[88] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[89] Yang ZL, Dai ZH, Yang YM, Carbonell J, Salakhutdinov R, Le QV. XLNet: Generalized autoregressive pretraining for language understanding. In: Proc. of the 33rd Conf. on Neural Information Processing Systems. Vancouver, 2019. 5754–5764.

[90] Zhang ZK, Pang WG, Xie WJ, Lv MS, Wang Y. Deep learning for real-time applications: A survey. Ruan Jian Xue Bao/Journal of

Software, 2020, 31(9): 2654–2677. (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5946.htm [doi: 10.13328/j.cnki.jos.005946]

[91] Wei W, Wu JS, Zhu CS. Special issue on deep learning for natural language processing. Computing, 2020, 102(3): 601–603. [doi: 10.1007/s00607-019-00788-3]

[92] Li FL, Ke J. Text representation method based on deep learning. Information Science, 2019, 37(1): 156–164 (in Chinese with English abstract). [doi: 10.13833/j.issn.1007-7634.2019.01.024]

[93] Malte A, Ratadiya P. Evolution of transfer learning in natural language processing. arXiv: 1910.07370, 2019.

[94] Yilmaz S, Toklu S. A deep learning analysis on question classification task using Word2Vec representations. Neural Computing and Applications, 2020, 32(7): 2909–2928. [doi: 10.1007/s00521-020-04725-w]

[95] Hewitt J, Manning CD. A structural probe for finding syntax in word representations. In: Proc. of NAACL-HLT. Minneapolis: Association for Computational Linguistics, 2019. 4129–4138.

[96] Ltaifa IB, Hlaoua L, Romdhane LB. Hybrid deep neural network-based text representation model to improve microblog retrieval. Cybernetics and Systems, 2020, 51(2): 115–139. [doi: 10.1080/01969722.2019.1705548]

[97] Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks, 2015, 61: 85–117. [doi: 10.1016/j.neunet.2014.09.003]

[98] Bengio Y, Lamblin O, Popovici D, Larochelle H. Greedy layer-wise training of deep networks. In: Proc. of the 19th Int'l Conf. on Neural Information Processing Systems. Cambridge: MIT Press, 2006. 153–160.

[99] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436–444. [doi: 10.1038/nature14539]

[100] Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. IEEE Trans. on Neural Networks, 2009, 20(1): 61–80. [doi: 10.1109/TNN.2008.2005605]

[101] Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: A survey. Knowledge-based systems, 2018, 151: 78–94. [doi: 10.1016/j.knosys.2018.03.022]

[102] Niepert M, Ahmed M, Kutzkov K. Learning convolutional neural networks for graphs. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York, 2016. 2014–2023.

[103] Gui J, Sun ZN, Wen YG, Tao DC, Ye JP. A review on generative adversarial networks: Algorithms, theory, and applications. Journal of Latex Classifiers, 2015, 14(8): 1–28.

[104] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. In: Advances in Neural Information Processing Systems. Montreal, 2014. 2672–2680.

[105] Zhang XM, Zhu XB, Zhang XY, Zhang NG, Li P, Wang L. SegGAN: Semantic segmentation with generative adversarial network. In: Proc. of the 4th IEEE Int'l Conf. on Multimedia Big Data (BigMM). Xi'an: IEEE, 2018. 1–5. [doi: 10.1109/BigMM.2018.8499105]

[106] Wang HW, Wang JL, Wang JL, Zhao M, Zhang WN, Zhang FZ, Xie X, Guo MY. GraphGAN: Graph representation learning with generative adversarial nets. IEEE Trans. on Knowledge and Data Engineering, 2017: 99–117.

[107] Wang C, Deng CY, Vladimir I. SAG-VAE: End-to-end joint inference of data representations and feature relations. In: Proc. of the 2020 Int'l Joint Conf. on Neural Networks (IJCNN). Glasgow: IEEE, 2020. 1–9. [doi: 10.1109/IJCNN48605.2020.9207154]

[108] Iancu B, Ruiz L, Ribeiro A, Isufi E. Graph-adaptive activation functions for graph neural networks. In: Proc. of 30th IEEE Int'l Workshop on Machine Learning for Signal Processing (MLSP). Espoo: IEEE, 2020. 1–6. [doi: 10.1109/MLSP49062.2020.9231732]

[109] Lu RK, Liu JW, Wang YF, Xie HJ, Zuo X. Auto-encoder based co-training multi-view representation learning. In: Proc. of the 23rd Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining. Macau: Springer, 2019. 119–130. [doi: 10.1007/978-3-030-16142-2_10].

[110] Vlachostergiou A, Caridakis G, Mylonas P, Stafylopatis A. Learning representations of natural language texts with generative adversarial networks at document, sentence, and aspect level. Algorithms, 2018, 11(10): 164. [doi: 10.3390/a11100164]

[111] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Bangkok: MIT Press, 2014. 3104–3112.

[112] Zhu X, Hu JT, Song LC, Suo GL, Zhan Y. Attention-based encoder-decoder model for photovoltaic power generation prediction. Journal of Physics: Conference Series, 2020, 1575: 012025. [doi: 10.1088/1742-6596/1575/1/012025]

[113] Nie YP, Han Y, Huang JM, Jiao B, Li AP. Attention-based encoder-decoder model for answer selection in question answering. Frontiers of Information Technology & Electronic Engineering, 2017, 18(4): 535–544. [doi: 10.1631/FITEE.1601232]

[114] Tian T, Fang Z. Attention-based autoencoder topic model for short texts. Procedia Computer Science, 2019, 151: 1134–1139. [doi: 10.1016/j.procs.2019.04.161]

[115] Liang ZQ, Pan D, Xu RJ. Knowledge representation framework of accounting event in corpus-based financial report text. Cluster Computing, 2019, 22(4): 9335–9346. [doi: 10.1007/s10586-018-2153-8]

[116] Wang XC, Liu ZT. Event semantic representation for news texts. Journal of Shanghai University (Natural Science), 2019, 25(5): 733–741 (in Chinese with English abstract). [doi: 10.12066/j.issn.1007-2861.1989]

[117] Giallonardo E, Poggi F, Rossi D, Zimeo E. Semantics-driven programming of self-adaptive reactive systems. Int'l Journal of Software Engineering and Knowledge Engineering, 2020, 30(6): 805–834. [doi: 10.1142/S0218194020400082]

[118] Wang XL, Feng A, Golshan B, Halevy A, Mihaila G, Oiwa H, Tan WC. Scalable semantic querying of text. Proc. of the VLDB Endowment, 2018, 11(9): 961–974. [doi: 10.14778/3213880.3213887]

[119] Liu FL, Liu YX, Ren XC, He XD, Sun X. Aligning visual regions and textual concepts for semantic-grounded image representations. In: Proc. of the 33rd Conf. on Neural Information Processing Systems (NeurIPS). Vancouver, 2019. 6847–6857.

[120] Liu J, Yang YH, He HH. Multi-level semantic representation enhancement network for relationship extraction. Neurocomputing, 2020, 403: 282–293. [doi: 10.1016/j.neucom.2020.04.056]

[121] Franco-Salvador M. A cross-domain and cross-language knowledge-based representation of text and its meaning. Procesamiento del Lenguaje Natural, 2019, (62): 111–114.

[122] Tang X, Chen L, Cui J, Wei BG. Knowledge representation learning with entity descriptions, hierarchical types, and textual relations. Information Processing & Management, 2019, 56(3): 809–822. [doi: 10.1016/j.ipm.2019.01.005]

[123] Liu CF, Zhang Y, Yu M, Li XW, Zhao MK, Xu TY, Yu J, Yu RG. Text-enhanced knowledge representation learning based on gated convolutional networks. In: Proc. of the 31st Int'l Conf. on Tools with Artificial Intelligence (ICTAI). Portland: IEEE, 2019. 308–315. [doi: 10.1109/ICTAI.2019.00051]

[124] Piad-Morffis A, Muñoz R, Almeida-Cruz Y, Gutiérrez Y, Estevez-Velarde D, Montoyo A. A neural network component for knowledge-based semantic representations of text. In: Proc. of the Recent Advances in Natural Language Processing. Varna, 2019. 904–911. [doi: 10.26615/978-954-452-056-4_105]

[125] Li HR, Zhu JN, Ma C, Zhang JJ, Zong CQ. Read, watch, listen, and summarize: Multi-modal summarization for asynchronous text, image, audio and video. IEEE Trans. on Knowledge and Data Engineering, 2019, 31(5): 996–1009. [doi: 10.1109/TKDE.2018.2848260]

[126] Li S, Tao ZQ, Li K, Fu Y. Visual to text: Survey of image and video captioning. IEEE Trans. on Emerging Topics in Computational Intelligence, 2019, 3(4): 297–312. [doi: 10.1109/TETCI.2019.2892755]

[127] Ibrahim ZAA, Saab M, Sbeity I. VideoToVecs: A new video representation based on deep learning techniques for video classification and clustering. SN Applied Sciences, 2019, 1(6): 560. [doi: 10.1007/s42452-019-0573-6]

[128] Yu LY, Yu J, Ling Q. Deep neural network based 3D articulatory movement prediction using both text and audio inputs. In: Proc. of the 2019 Int'l Conf. on Multimedia Modeling. Thessaloniki: Springer, 2019. 68–79. [doi: 10.1007/978-3-030-05710-7_6]

[129] Han ZZ, Shang MY, Wang XY, Liu YS, Zwicker M. Y2Seq2Seq: Cross-modal representation learning for 3D shape and text by joint reconstruction and prediction of view and word sequences. Proc. of the AAAI Conf. on Artificial Intelligence, 2018, 33(1): 126–133. [doi: 10.1609/aaai.v33i01.3301126]

[130] Suwela N. Ranking index berita new normal dengan metode information retrieval menggunakan vector space model. STRING (Satuan Tulisan Riset dan Inovasi Teknologi), 2020, 5(1): 61–69. [doi: 10.30998/string.v5i1.6479]

[131] Zhang T, Shen S, Cheng CX, Su K, Zhang XX. A topic model based framework for identifying the distribution of demand for relief supplies using social media data. Int'l Journal of Geographical Information Science, 2021, (10): 1–22. [doi: 10.1080/13658816.2020.1869746]

[132] Shah SMA, Ge HW, Haider SA, Irshad M, Noman SM, Arshad J, Ahmad A, Younas T. A quantum spatial graph convolutional network for text classification. Computer Systems Science and Engineering, 2021, 36(2): 369–382. [doi: 10.32604/csse.2021.014234]

[133] Jiang Z, Gao S. Study on text representation method based on deep learning and topic information. Journal of Computing, 2020, 102(3): 623–642. [doi: 10.1007/s00607-019-00755-y]

[134] Pan J, Wu ZD. A review of word representation learning. Journal of the China Society for Scientific and Technical Information, 2019, 38 (11): 1222–1240 (in Chinese with English abstract). [doi: 10.3772/j.issn.1000-0135.2019.11.010]

[135] Sun YM, Lin L, Tang DY, Yang N, Ji ZZ, Wang XL. Modeling mention, context and entity with neural networks for entity disambiguation. In: Proc. of the 24th Int'l Conf. on Artificial Intelligence (IJCAI). Buenos Aires: AAAI Press, 2015. 1333–1339.

[136] Rezaeinia SM, Rahmani R, Ghodsi A, Veisi H. Sentiment analysis based on improved pre-trained word embeddings. Expert Systems with Applications, 2019, 117: 139–147. [doi: 10.1016/j.eswa.2018.08.044]

[137] Wang Y, Sun YN, Ma ZC, Gao LS, Xu Y. Named entity recognition in Chinese medical literature using pretraining models. Scientific Programming, 2020, 2020: 8812754. (in Chinese with English abstract). [doi: 10.1155/2020/8812754]

[138] Bhattacharjee A, Hasan T, Samin K, Rahman MS, Iqbal A, Shahriyar R. BanglaBERT: Combating embedding barrier for low-resource language understanding. arXiv: 2101.00204, 2021.

附中文参考文献:

[4] 于剑. 语言与图灵测试. 自动化学报, 2016, 42(5): 668–669. [doi: 10.16383/j.aas.2016.y000004]

[6] 梁君英, 刘海涛. 语言学的交叉学科研究: 语言普遍性、人类认知、大数据. 浙江大学学报(人文社会科学版), 2016, 46(1): 108–118. [doi: 10.3785/j.issn.1008-942X.CN33-6000/C.2015.10.231]

[8] 孙茂松, 刘挺, 姬东鸿, 穗志方, 赵军, 张钹, 吾守尔·斯拉木, 俞士汶, 朱军, 李建民, 刘洋, 王厚峰, 吐尔根·依布拉音, 刘群, 刘知远. 语言计算的重要国际前沿. 中文信息学报, 2014, 28(1): 1–8. [doi: 10.3969/j.issn.1003-0077.2014.01.001]

[9] 宗成庆. 统计自然语言处理. 第2版, 北京: 清华大学出版社, 2013. 8.

[11] 褚晓敏, 朱巧明, 周国栋. 自然语言处理中的篇章主次关系研究. 计算机学报, 2017, 40(4): 842–860. [doi: 10.11897/SP.J.1016.2017.00842]

[12] 李仕春. 论世界语言学学术思想变迁之大势. 东岳论丛, 2017, 38(8): 163–168. [doi: 10.15981/j.cnki.dongyueluncong.2017.08.022]

[13] 赵永刚. 语言的进化与生物语言学进路诠疏——兼评《为什么只有我们: 语言与进化》. 学术探索, 2018, (6): 107–116. [doi: 10.3969/j.issn.1006-723X.2018.06.016]

[14] 冯志伟. 自然语言处理的形式模型. 合肥: 中国科学技术大学出版社, 2010. 1.

[15] 张磊, 卫乃兴. 局部语法的演进、现状与前景. 当代语言学, 2018, 20(1): 103–116.

[38] 徐通锵. “字本位”和语言研究. 语言教学与研究, 2005, (6): 1–11.

[39] 张若男. 字本位理论视角下的对外汉语教学研究述评. 现代语文, 2018, (4): 134–139.

[46] 邱锡鹏. 神经网络与深度学习. 北京: 机械工业出版社, 2020. 4.

[49] 董蕊芳, 柳长安, 杨国田. 一种基于改进TF-IDF的SLAM回环检测算法. 东南大学学报(自然科学版), 2019, 49(2): 251–258. [doi: 10.3969/j.issn.1001-0505.2019.02.008]

[60] 赵京胜, 张丽, 朱巧明, 周国栋. 中文文学作品中的社会网络抽取与分析. 中文信息学报, 2017, 31(2): 99–106, 116.

[63] 赵京胜, 朱巧明, 周国栋, 张丽. 自动关键词抽取研究综述. 软件学报, 2017, 28(9): 2431–2449. http://www.jos.org.cn/1000-9825/5301.htm [doi: 10.13328/j.cnki.jos.005301]

[90] 张政馗, 庞为光, 谢文静, 吕鸣松, 王义. 面向实时应用的深度学习研究综述. 软件学报, 2020, 31(9): 2654–2677. http://www.jos.org.cn/1000-9825/5946.htm [doi: 10.13328/j.cnki.jos.005946]

[92] 李枫林, 柯佳. 基于深度学习的文本表示方法. 情报科学, 2019, 37(1): 156–164. [doi: 10.13833/j.issn.1007-7634.2019.01.024]

[116] 王先传, 刘宗田. 新闻文本中事件语义表示. 上海大学学报(自然科学版), 2019, 25(5): 733–741. [doi: 10.12066/j.issn.1007-2861.1989]

[134] 潘俊, 吴宗大. 词汇表示学习研究进展. 情报学报, 2019, 38(11): 1222–1240. [doi: 10.3772/j.issn.1000-0135.2019.11.010]

赵京胜(1969－), 男, 副教授, 主要研究领域为自然语言处理, 中文信息处理.

高祥(1992－), 男, 硕士, 主要研究领域为智能信息处理.

宋梦雪(1996－), 女, 硕士, 主要研究领域为智能信息处理.

朱巧明(1963－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为自然语言处理, 智能信息处理.