中国科学: 信息科学 2018 年 第 48 卷 第 11 期: 1467–1486
SCIENTIA SINICA Informationis
社会媒体计算与自然语言处理专刊 • 评述

量子语言模型研究综述

张鹏[1*], 马鑫典[1], 宋大为[2]

1. 天津大学计算机科学与技术学院, 天津 300350
2. 北京理工大学计算机学院, 北京 100081
* 通信作者. E-mail: pzhang@tju.edu.cn

收稿日期: 2018–06–22; 接受日期: 2018–09–10; 网络出版日期: 2018–11–09

国家重点研发计划 (批准号: 2017YFE0111900) 和国家自然科学基金 (批准号: U1636203, 61772363) 资助项目

摘要 语言模型是自然语言处理相关领域研究工作的重要基础. 近年来, 人们基于量子力学概率理论提出量子语言模型. 本文旨在综述量子语言模型的研究动机和当前进展. 我们首先回顾语言模型的研究现状及存在的问题, 然后介绍信息检索领域和语音处理领域的量子语言模型, 以及我们所提出的应用于自动问答领域的端到端的量子语言模型. 通过分析各种量子语言模型的优缺点, 以及量子力学与神经网络的本质联系, 提出进一步的研究思路与未来愿景.

关键词 语言模型, 量子语言模型, 神经网络, 信息检索, 量子力学

1 语言模型的研究现状

语言作为社会文明发展与传递的主要媒介之一, 对社会、政治、科技及文化等各方面产生深远且重要的影响. 现今时代, 随着计算机科学的飞速发展以及人工智能的长足进步, 作为一个核心研究方向, 自然语言处理的重要性不言而喻. 自然语言处理的基本任务是利用计算机对自然语言 (即人类语言) 的内在规律进行建模, 从而进一步进行自然语言的生成与理解. 因此, 语言模型的研究进展是重中之重. 发展高效而鲁棒的语言模型有助于支持诸如搜索引擎、智能对话、在线推荐及电子商务等实际应用, 不仅具有重要的理论意义, 同样蕴含巨大的社会价值.

语言模型作为自然语言处理的核心问题, 在取得了一系列发展的同时, 也暴露出诸多问题. 起初, 科学家们根据语言学知识和领域知识, 人工编制一些文法规则, 设计出文法型语言模型 [1,2]. 但是, 这种语言模型 [2] 因为人工成本高且只针对特定领域, 不能处理大规模通用领域的文本. 于是, 基于统计不确定性的语言模型应运而生, 计算机通过估计统计语言模型的概率分布参数, 推断自然语言片段出现的可能性. 时至今日, 统计语言模型已经发展出很多具体的模型 [3], 每种模型都有其基本假设, 这些基本假设在具体化语言模型的同时, 也限制了其在某些方面的建模能力, 例如建模词与词之间的依赖关系 [3]、语义片段的潜在语义 [4] 以及隐藏意图等方面.

引用格式: 张鹏, 马鑫典, 宋大为. 量子语言模型研究综述. 中国科学: 信息科学, 2018, 48: 1467–1486, doi: 10.1360/N112018-00163
Zhang P, Ma X D, Song D W. A survey of quantum language models (in Chinese). Sci Sin Inform, 2018, 48: 1467–1486, doi: 10.1360/N112018-00163

具体而言, 词袋模型 (bag of words) [5] (即一元语言模型) 假设将一篇文档或者一个词的序列看成是词的集合, 并且集合中的词是无序和独立的. 尽管词袋模型在一些应用任务 (如 ad-hoc 信息检索 [6]) 取得了不错的效果, 但是无法表示词与词之间的语义关联. 针对此问题, 多元语言模型 (n-gram) [7] 建模当前词与其前面 $n-1$ 个词的相关性, 通过考虑与当前词相邻近的词来估计语言模型的概率分布, 但是这种方法增加了概率参数估计规模, 另外其更本质的局限性在于有些强语义关联的词汇在局域范围内并不是共现的, 基于局域物理位置相似性 (proximity-based) 的方法不能很好地刻画潜在的或全局的语义关联性.

面对词袋模型与多元语言模型的不足, 潜在语义索引模型 (latent semantic index, LSI) [8] 首先将每个词嵌入到一个由奇异值分解 (SVD) 得到的特征向量张成的子空间中, 并假设在这个 "潜在的" 子空间中, 语义关联的词项之间也是相似的 (例如其向量内积所表示的余弦相似度较小). 通过这个假设, LSI 试图刻画多义词和近义词等全局语义关联. 类似地, 神经语言模型 [9,10] 的方法也基于上述子空间嵌入假设 [11], 所不同的是其训练模型是神经网络模型. 潜在语义索引模型和神经语言模型都试图找到能够刻画词与词之间语义相似性的子空间, 但是语义相似性其实并不必然等同于实际应用所需要的语义性质, 例如在信息检索中, 相关性 (relevance) 就不必然等同于相似性 (similarity).

为更好地建模语义关联, 信息检索领域的学者们提出利用 Markov 随机场 (Markov random field, MRF) 改进语言模型 [12]. 在估计文档的相关性时, MRF 模型按照查询词不同的组合 (比如单个词、多个词组成的词组等) 分别算出文档的相关性分数, 然后计算总的相关性分数. 但是该模型在计算文档相关性分数时, 只是将不同的依赖 (或特征) 信息得到的分数做线性加权, 并不能有机地将不同语义依赖统一到一种原则性的语言模型框架之下 [13].

同样针对这一问题, Sordoni 等 [13] 提出一种量子语言模型 (quantum language model, QLM), 试图利用量子力学的密度矩阵建模文本序列 (例如查询词和文档) 中词与词之间的依赖关系. 简单地说, 在量子语言模型中, 投影算子 (projector) 表示单个词或词的组合 (可以看作是基准的量子态), 密度矩阵可以用来测量各个可观测 (observed) 量子态 (例如查询词) 出现的概率, 查询和文档所对应的密度矩阵可以用极大似然估计方法求得, 然后利用两个密度矩阵的 Von-Neumann 散度 (VN-divergence) 计算查询和文档的相关性. 实验表明在 ad-hoc 信息检索中, 相比一元语言模型和基于 MRF 的高阶语言模型, 量子语言模型取得显著的性能提升.

尽管如此, 当前的量子语言模型研究工作仍然存在若干关键性的问题, 这些问题在一定程度上限制了量子语言模型的深入研究和广泛应用. 深入研究方面, 量子语言模型的局限之处在于: 首先, 虽然量子语言模型通过训练密度矩阵建模语义依赖 [13], 但从实质上而言, 这种语义依赖只对应查询词组成的词组, 未能充分考虑全局依赖. 其次, 量子语言模型只是应用于 ad-hoc 信息检索任务中, 该检索任务非常简单: 用户给定查询词, 然后系统给出检索结果, 尚未拓展到文本生成和自动对话的典型任务 [14] 上. 最后, 如何应对在大规模数据集下, 密度矩阵高效而快速的训练, 这些都是亟待研究的问题.

2 量子语言模型的研究背景

为了更好地了解量子语言模型的背景, 有必要从建模词与词之间的依赖关系、语义关联性入手, 系统地回顾统计语言模型, 以及信息检索领域的语言模型. 因为量子语言模型是在信息检索领域首先提出的, 我们也得介绍量子力学在信息检索领域的发展历史.

2.1 统计语言模型

统计语言模型 [6] 通过表示语言基本单位 (例如词、词组、句子等, 这里统一叫做文本片段) 的分布函数, 力图描述语言的统计生成规则. 但是因为自然语言单词量大、句法复杂, 很难计算句子的概率空间, 所以一般是根据某些假设条件, 将计算文本片段的的概率分解为边缘概率或者条件概率的乘积. 例如在 n-gram 模型中, 采用了 Markov 假设, 认为每个单词只与其前面 $n-1$ 个词 (上下文) 相关, 参数 n 成为模型的阶数. 一阶语言模型 (又称为一元语言模型), 假设词与词之间是统计独立的, 随着阶数的增大, 模型的复杂性不断提高. n-gram 模型成功捕捉了自然语言局部依赖的性质 (例如上下文相邻词的依赖关系), 但却不能表示语言中存在的远程依赖 (例如句子结构、语义关系等).

决策树语言模型 [15,16] 的提出, 解决了 n-gram 模型的复杂性及冗余上下文依赖的问题. 具体而言, 决策树语言模型通过设计一些问题或条件, 尽量保留与当前预测变量 (例如当前词) 关联的上下文, 去掉不关联的上下文. 随着决策树高度的增加, 每个节点的训练语料相对的就会减少, 导致精度的降低, 产生所谓的数据碎化问题. 于是, 科学家们提出最大熵模型 [17], 基本思想是将统计语言模型的估计问题看作是有约束的概率分布优化问题. 总体来说, 决策树语言模型和最大熵模型在解决 n-gram 模型上下文只用模型阶数约束的同时, 也存在约束条件设计麻烦以及时间复杂性高等问题. 这些问题在一定程度上, 限制了其在大规模文本处理任务 (例如信息检索) 中的广泛使用.

2.2 经典信息检索

信息检索 (information retrieval, IR) 是自然语言处理领域的一个重要研究方向, 其典型的应用是通用搜索引擎, 例如百度、Google 等 [18]. 信息检索是对于用户的信息需求 (通常用查询词来表示), 找出与之相关的信息载体 (通常由文档来表示). 信息检索模型是在表示信息需求和信息载体的同时, 计算信息载体相对于信息需求的相关程度. 我们主要介绍基于概率统计的检索模型或语言模型在信息检索中的应用, 以及它们建模语义依赖的作用.

在概率检索模型和统计语言模型之前, 科学家们已提出其他一些经典信息检索模型, 例如布尔模型和空间向量模型 [19] 等. 布尔模型 [20] 是最早的一种检索模型, 它将查询表示为布尔表达式, 然后输出符合布尔表达式的文档. 这种检索方式将相关性看成是一个二值属性 (即相关和不相关两种属性), 虽然在一些图书管理系统中得到成功应用, 但对于通用搜索环境来说, 无法针对信息相关程度排序. 向量空间模型 (vector space model) [21] 的基本原理是将查询和文档分别表示成一个向量, 向量中每一个索引项的值可以是 TF/IDF 值 [22], 表示每个词的权重, 然后通过计算查询和文档之间向量的内积或者余弦相似度的方式, 求解相关性得分. 这种方法可以对文档进行排序, 并且改善了检索性能.

概率检索模型 (probabilistic retrieval model) [23] 首次将概率统计的不确定性引入检索模型, 这是检索过程具有不确定性的体现. 通过一些基本假设 (例如词与词之间是独立的、索引项权重和文档相关性是二值的), 利用 Bayes 公式推导出文档与查询的相关性计算公式. 这种方法有概率统计的理论基础, 例如在符合基本假设的前提下, 证明概率检索模型可以提供最优的检索结果 [23]. 从实验角度, 概率检索模型表现出优秀的检索性能, 例如其导出的 BM25 模型仍是现在搜索系统的一个很有效的特征 [24].

语言模型也是一种基于概率统计的检索模型, 与概率检索模型的不同之处在于其概率生成的原则 [25]: 概率检索模型假设查询 Q 生成相关文档集 D, 而语言模型假设文档集 D 生成查询 Q. 因此, 语言模型的核心问题是计算语言模型生成查询的概率 (表示为 $p(Q|D)$, 称为查询似然性). 通过 Bayes 法则和词间独立性假设, $p(Q|D)$ 就可以很容易的由文档语言模型的边缘概率的乘积求得. 语言模型形

式简单且有较好的理论基础, 在对文档做平滑 (smoothing) 之后, 其计算公式可以分解出 3 个信息检索典型的特征, 即词频 (term frequency, TF)、逆向文档频率 (inverse document frequency, IDF) 以及文本长度 (document length, DL) [26].

可以看出, 虽然上述基于统计的检索模型各有侧重, 但有一个共同的假设, 即不同词语之间是统计独立的, 无法建模词语之间的语义关联. 为更好地刻画语义关联 (特别是不同查询词之间的语义关联), 研究者们提出利用 Markov 随机场 (Markov random field, MRF) 改进语言模型 [12]. 在估计文档的相关性时, MRF 模型按照查询词不同的组合 (比如单个词、多个词组成的词组等) 分别计算文档的相关性分数, 然后利用线性拟合方法将这些分数集成起来形成总的相关性分数. 为了计算语义分布相似度, 建模语义关联, 研究者们提出翻译语言模型 [27], 通过融合多个训练集的翻译知识, 以提升领域翻译知识的权重来建模语义关联. 另外, 一些其他词聚类方法, 例如 Latent Dirichlet Allocation (LDA) [28] 和 Partially Observable Markov Decision Processes (POMDP) [29] 等, 也被用于改进语言模型. 但是上述工作在计算文档相关性分数时, 只是将不同的依赖信息 (或特征) 得到的分数做线性加权, 并不能有机地将不同的依赖信息统一到一种原则性的 (principled) 语言模型框架之下 [13].

2.3 量子信息检索

首先, 需要澄清的是, 所谓量子信息检索指借助量子理论 (quantum theory, QT) 的数学方法、实验直觉, 以及类量子现象[1)](例如宏观现象中体现出来的类量子规律) 来解决信息检索问题的研究工作. 也就是说, 量子信息检索并不是说在量子计算机上运行的信息检索, 或者说物理上必须建模量子级别的微观粒子. 量子信息检索起初是希望经典信息检索的模型更加一般化, 从而建模一些非经典的概率现象或特征.

van Rijsbergen 在其文献 [30] 中开创性地提出将传统信息检索模型 (例如布尔模型、向量空间模型和概率检索模型等) 统一在 Hilbert 向量空间中的量子力学形式化框架中. 具体来说, 量子理论为信息检索基本元素 (例如查询, 文档和多媒体信息等) 提供了在 Hilbert 复数空间中的几何表示. 在 van Rijsbergen 的启发下, 涌现出一些量子信息检索的工作.

首先, 在信息检索领域中, 人们开始探索和建模宏观类量子现象. 受量子干涉现象的启发, 研究了认知干涉, 即用户的先期阅读经历是否会改变其对当前文本的相关性判断 [31], 并研究了查询词的次序效应 (order effect), 以及相对应的量子干涉现象 [32]. Zuccon 等 [33] 指出了信息检索 (IR) 中的文本排序场景和量子理论中的双缝实验解释之间的同构性, 并主张在测度文本相关性时考虑文本之间的干涉. Sordoni 等 [34] 类比了量子力学中的双缝干涉实验, 将任意两个隐主题类比为双缝, 将主题词分布看作屏幕, 研究两个隐主题之间的干涉效应. 此外, 为了捕获搜索会话中的动态信息需求, 利用密度矩阵 [35] 的演化过程建立了自适应量子语言模型, 研究 Session Search 中的查询词的不确定性. 在建模宏观类量子现象方面, Zhang 等 [35~37] 提出了光子极化实验在信息检索排序模型和查询扩展模型的对应关系. 文献 [38] 提出利用纯相关依赖关系建模后测量设置下的量子纠缠, 抽取一些依赖关系更强的词组作为量子基本事件. 并在量子语言模型中建模了量子纠缠这一宏观类量子现象.

此外, 研究人员提出了若干基于 Hilbert 空间的信息检索模型和框架. 例如, 可以将用户的信息需求和信息对象表示为对应的子空间, 并融合不同维度的上下文信息 (例如, 文本、任务、用户或地点等). Piwowarski 等 [39] 利用张量空间与状态向量空间构造量子信息检索方法, 随后 Frommholz [40] 基于信息需求的多元表示扩展了上述框架, 为各个表示定义合适的 Hilbert 子空间. Sordoni 等 [13] 在量

1) 在宏观层面, 我们用类量子现象表述, 这有区别于微观的量子现象.

子概率框架下扩展了传统语言模型, 并提出了量子语言模型.

目前, 量子语言模型在信息检索主流研究中未能取得广泛应用, 原因有以下几点. 第一, 大多数同行认为量子力学主要针对微观世界, 与计算机的联系仅是量子计算. 但其实量子力学本身是一个数学框架, 已经应用于一些诸如社会学、经济学和认知科学等宏观领域 [41, 42], 并且其研究不依赖于量子计算. 第二, 在某种意义上说, 信息检索是一门实验科学, 而早期的量子信息检索模型在实验效果上未能表现出明显的优势. 第三, 量子语言模型提出来之后, 因为其密度矩阵计算成本较大, 加之不能利用监督信息, 所以它在很多任务上 (例如自动问答任务) 表现不佳. 这些原因都限制了量子信息检索和量子语言模型在主流研究中的影响力. 针对这些问题, 本文旨在澄清量子力学的数学框架及其在信息检索相关领域的研究动机, 并逐步叙述各种量子语言模型. 具体而言, 我们在本文第 3 节, 详细介绍量子力学公理假设, 及其与信息检索和自然语言处理任务的关系. 在第 4 节, 详细介绍 3 个不同领域的量子语言模型.

3 量子力学公理及应用实例

本节介绍量子力学的公理体系 [43] 及其与语言建模的联系. 20 世纪 30 年代数学家 von Neumann [44] 将量子理论进行公理化, 其概率测量是基于空间投影的测量方法, 这一形式化体系并不神秘, 主要是基于 Hilbert 空间下的线性代数和投影理论. 我们将介绍 4 个基本量子力学公理, 并且说明其在自然语言处理方面的应用实例.

3.1 状态空间

公理1 (量子叠加态) 假设一个量子比特有一个二维的状态空间, 用 $|0\rangle = (1,0)^{\mathrm{T}}$ 和 $|1\rangle = (0,1)^{\mathrm{T}}$ 构成这个空间的标准正交基, 则状态空间中的任意状态可用叠加态表示, 如式 (1) 所示:

$$|\phi\rangle = \alpha|0\rangle + \beta|1\rangle, \tag{1}$$

其中, α 和 β 是限定在实数域的概率振幅, $|\alpha|^2$ 和 $|\beta|^2$ 表示概率, 所以需要归一化, 即满足 $|\alpha|^2+|\beta|^2 = 1$.

在信息检索中, 用户对文档的相关性判断 (relevance judgement) 可以用量子叠加态表示 [12], $|d\rangle = \alpha|r\rangle + \beta|\neg r\rangle$, 其中, $|r\rangle$ 表示相关, $|\neg r\rangle$ 表示不相关, $|\alpha|^2$ 表示文档相关的概率值, $|\beta|^2$ 表示文档不相关的概率值. 另外, 一个词的多种词义也可以用量子叠加态表示 [13], 比如 “苹果” 这个词, 它既可以是水果, 也可以是手机. 故此, 我们使用一组完备的基向量 $(|c_1\rangle, |c_2\rangle, \ldots, |c_m\rangle)$ 表示一个词的多种词义, 如式 (2) 所示:

$$|\phi\rangle = \alpha_1|c_1\rangle + \alpha_2|c_2\rangle + \cdots + \alpha_m|c_m\rangle. \tag{2}$$

根据向量表示方法, 词向量 (word embedding) 就可以看作是一种叠加态的表示.

单个射线或向量 (量子叠加态) 对应纯态, 多个向量对应混合态. 混合态需要用密度矩阵表示. 在量子力学中, 一个系统的状态经常用 Hilbert 空间上的密度矩阵 ρ 表示. 密度矩阵需要满足两个条件: (1) 半正定; (2) 迹为 1, 即 $\mathrm{tr}(\rho) = 1$. 根据密度矩阵空间特征分布, 可以分为混合态的密度矩阵与纯态的密度矩阵. 可以通过式 (3) 判断密度矩阵的状态为纯态或混合态. 如果密度矩阵 ρ 的平方的迹等于 1, 则 ρ 是纯态, 小于 1 是混合态.

$$\mathrm{tr}(\rho^2)\begin{cases} = 1, & \text{pure state;} \\ < 1, & \text{mixed state.} \end{cases} \tag{3}$$

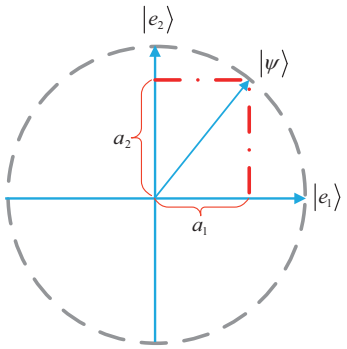


图 1 (网络版彩图) 投影测量的二维几何表示

Figure 1 (Color online) 2-dimensional geometric representation of projection measurement

直观理解就是, 如果密度矩阵是纯态, 可以用一个向量的外积运算得到矩阵, 即 $\rho = |\psi\rangle\langle\psi|$, 其中 $|\psi\rangle$ 是一个叠加态. 如果密度矩阵是混合态, 可以用多个向量外积运算再加权求和得到的矩阵. 即 $\rho = \sum_i p_i|\psi_i\rangle\langle\psi_i|$, 其中 $\sum_i p_i = 1$. 信息检索中, 用户在对文档的相关性判断时, 判断相关还是不相关, 是具有上下文性的, 需要借助量子测量得到. 通过量子测量, 量子的叠加状态会以一定的概率塌缩到文档相关和不相关这两种状态的其中一种. 词的含义也有上下文性, 一个给定的多义词通过量子测量会塌缩到一种具体的词义. 第 3.2 小节将具体介绍量子测量公理.

3.2 量子测量

量子力学中有很多重要的测量方法, 包括一般测量、投影测量和 POVM 测量等. 在量子语言模型建模过程中, 使用较多的一种测量方法是投影测量.

公理2 (投影测量) 有一个可观测系统的状态, 由 Hilbert 空间向量 $|\psi\rangle = \sum_{i=1}^{n} a_i|e_i\rangle$ 表示, 用 $\Pi_i = |e_i\rangle\langle e_i|$ 表示本征空间上的一个测量算子. 测量状态 $|\psi\rangle$ 时, 得到概率 $P(e_i|\psi)$:

$$P(e_i|\psi) = \langle\psi|\Pi_i|\psi\rangle = \langle e_i|\psi\rangle^2 = a_i^2. \tag{4}$$

测量后的量子系统状态为

$$|\psi'\rangle = \frac{\Pi_i|\psi\rangle}{\sqrt{P(e_i|\psi)}}, \tag{5}$$

其中, a_i 是概率振幅, a_i^2 是概率, 满足 $\sum_i a_i^2 = 1$, $\langle e_i|\psi\rangle$ 表示向量的内积运算.

为了能够形象的表示测量过程, 投影测量的二维几何表示如图 1 所示, $|e_1\rangle$ 和 $|e_2\rangle$ 对应两个基本量子事件, 用两个基向量表示. 量子系统状态 $|\psi\rangle$ 是一个量子叠加态的表示, 即 $|\psi\rangle = a_1|e_1\rangle + a_2|e_2\rangle$, 朝两个不同的方向作投影, 可以分别得到概率 $p(e_1|\psi)$ 和 $p(e_2|\psi)$. $p(e_1|\psi) = a_1^2$ 表示该量子系统的状态塌缩到量子基本事件 e_1 的概率, $p(e_2|\psi) = a_2^2$ 表示该量子系统状态塌缩到量子基本事件 e_2 的概率.

投影测量对应向量的内积计算, 可以用来表示余弦相似度的度量. 在信息检索中, 假设有一个检索的查询 q 和一篇文档 d, 用一组向量来表示 q 和 d, q 和 d 之间的相似度公式 (6) 如下:

$$\cos^2(q, d) = |\langle q, d\rangle|^2 = p(q|d). \tag{6}$$

两个词向量的相似度也可以用余弦相似度来刻画.

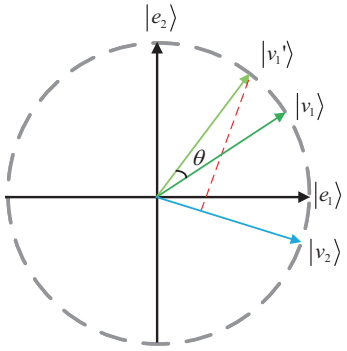


图 2 (网络版彩图) 酉演化二维空间示意图

Figure 2 (Color online) Unitary evolution 2-dimensional spatial diagram

3.3 酉演化

一个封闭的量子系统的演化可以由一个酉变换来刻画, 即系统的状态 $|\psi\rangle$ 是随时间变化的, 量子力学理论为这种量子系统的变化提供了一种方法, 即酉演化.

公理3 (酉演化) 系统在时刻 t_1 的状态 ψ 和系统在时刻 t_2 的状态 $|\psi'\rangle$, 可以通过一个 U 算子 (即酉矩阵) 进行状态变化.

$$|\psi'\rangle = U|\psi\rangle. \tag{7}$$

定义一个二维的实数域酉矩阵: $U = \left[\begin{smallmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{smallmatrix}\right]$, 一个二维空间状态, 用空间向量 $|\psi\rangle = [a_1, a_2]^{\mathrm{T}}$ 表示. 那么, 酉变化可以表示为

$$|\psi'\rangle = U|\psi\rangle = \begin{bmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}a_1 \\ a_2\end{bmatrix}. \tag{8}$$

在量子语言建模过程中, 酉演化扮演一个重要的角色, 它通常用在连续的测量过程中. 假设有一个文本序列, $S = (w_1, w_2, \ldots, w_n)$, 序列 S 中的每一个词可以用一个叠加态的向量表示. 在语言模型的建模过程中, 需要计算在当前词序情况下, 出现下一个词的条件概率, 一直循环计算, 到句子结束, 然后用这些概率值计算句子的困惑度 (perplexity), 这是一个 Markov 链的过程. Sordoni 等 [13] 在量子概率框架下提出的量子语言模型是计算单个词或词组的概率值, 不能够用来建模句子序列的条件概率. 为了建模句子的序列关系, 文献 [45] 用到了酉演化公理, 酉演化可以弥补量子测量建模句子序列的不足.

为了便于直观理解, 基于基向量 $|e_1\rangle$ 和 $|e_2\rangle$ 建立一个二维的空间示意图, 如图 2 所示. “买了” 对应空间状态向量 $|v_1\rangle$, “书” 对应空间状态向量 $|v_2\rangle$. 假设我们的句子状态空间向量已经对 “张三” 这个词做了投影测量, 测量后的状态就落在了状态 $|v_1\rangle$ 上, 也就是 “买了” 对应的空间状态, 如果不进行空间酉变化, 就会从 $|v_1\rangle$ 的状态开始向 $|v_2\rangle$ 状态做投影测量, 这样就不能很好地建模句子的整体语义. 对应语言模型上的理解就是, “书” 这个词出现的概率等于在 “买了” 这个词出现的条件下 “书” 出现的一元条件概率, “张三” 与 “书” 就是相对独立的, 无法表示在 “张三” 和 “买了” 出现的条件下 “书” 出现的条件概率, 即二元条件概率. 因此需要一个酉演化, 让状态空间变化到 $|v_1'\rangle$ 之后再向 $|v_2\rangle$ 作投影.

3.4 复合系统

公理4 一个复合系统的状态空间是由多个子系统的状态空间做张量积得到的, 若将多个子系统编号为 1 到 n, 子系统 i 的状态被置为 ψ_i, 则整个系统的总状态为 $|\psi_1\rangle \otimes |\psi_2\rangle \otimes \cdots \otimes |\psi_n\rangle$, 即复合系统的状态.

假设我们有两个子系统, 分别为 $|0\rangle = [1,0]^{\mathrm{T}}$ 与 $|1\rangle = [0,1]^{\mathrm{T}}$, 张量积之后为

$$|0\rangle \otimes |1\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \otimes \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1\times 0 \\ 1\times 1 \\ 0\times 0 \\ 0\times 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}. \tag{9}$$

在量子语言建模的过程中, 这里用建模一个句子序列为例, $S = (w_1, w_2, \ldots, w_n)$, 把每一个单词 w_i 作为一个子系统, 用向量 $|w_i\rangle$ 表示, 那么多个子系统的张量积就是一个句子序列 S 对应的复合系统 $|S\rangle$:

$$|S\rangle = |w_1\rangle \otimes |w_2\rangle \otimes \ldots \otimes |w_n\rangle.$$

为了便于理解, 依然以句子 “张三/买了/书” 为例. 每个单词用一个独热表示 (one-hot), 分别是 “张三” 对应 $|w_1\rangle = (1,0,0)^{\mathrm{T}}$, “买了” 对应 $|w_2\rangle = (0,1,0)^{\mathrm{T}}$, “书” 对应 $|w_3\rangle = (0,0,1)^{\mathrm{T}}$, 那么这个句子的复合系统的状态表示为 $|w_1\rangle \otimes |w_2\rangle \otimes |w_3\rangle$, 每个向量 $|w_i\rangle$ 也可以用词向量来表示.

4 量子语言模型的详细介绍

基于上述量子力学公理体系, 本节详细介绍信息检索领域的量子语言模型、语音处理领域的量子语言模型和自动问答领域的量子语言模型. 信息检索领域的量子语言模型, 主要研究短文本与长文本的匹配问题, 即查询词与文档之间的匹配, 计算检索词与文档之间的匹配分数, 并根据匹配分数进行文档排序. 信息检索领域的量子语言模型 [13] 主要使用了量子公理的 1 和 2, 分别从词的表示, 词与词依赖关系的表示, 以及文档的表示展开叙述, 然后介绍基于最大似然函数的密度矩阵优化, 以及文档和查询的匹配函数. 语音处理领域的量子语言模型, 主要研究使用量子测量的方法计算每个词在历史词出现条件下的条件概率, 然后进一步计算句子中所有词出现的联合概率, 从而去学习数据的分布情况. 语音处理领域的量子语言模型 [45] 使用了公理 2 和 3, 重点体现在量子公理 3, 即酉演化, 主要是建模句子序列, 计算词的条件概率与句子的联合概率. 然后, 介绍在自动问答领域的量子语言模型, 在该领域, 量子语言模型需要考虑问题与答案是否能够精确匹配, 研究的是短文本与短文之间的匹配问题, 需要获取更多的短文本语义信息. 在自动问答领域, 我们提出的基于神经网络的端到端量子语言模型 [46] 介绍了句子的密度矩阵表示, 问答句对的联合表示和两种更新联合矩阵的方法.

4.1 信息检索领域的量子语言模型

为了更好地建模语义关联, Sordoni 等 [13] 提出一种信息检索领域的量子语言模型, 这种语言模型利用量子概率的知识, 以及利用量子力学的密度矩阵建模文本序列 (例如查询词和文档) 中词与词之间的依赖关系, 这里体现了量子力学公理中的叠加态公理和测量公理. 简单地说, 在量子语言模型中, 投影算子 (projector) 表示单个词或词的组合, 密度矩阵可以用来测量各个可观测的 (observed) 量子态

(例如查询词) 出现的概率, 对应的密度矩阵 (查询与文档) 提出用极大似然估计的方法求得, 这里使用了一种 $R\rho R$ [47] 的方法, 实质是一种 Expectation maximization (EM) 算法. 然后利用两个密度矩阵的 VN-divergence 计算查询和文档的相关性, 利用相关性的大小进行排序. 接下来, 将详细介绍该语言模型.

4.1.1 单个词的表示

在使用经典概率建模语言时, 样本空间是一个集合, 而使用量子概率建模语言时, 样本空间是一个 Hilbert 空间. 因此每个词的表示可以由其对应的 Hilbert 空间的空间向量做外积得到的投影算子表示. 这个过程可以理解为单个词的向量表示到量子事件空间的映射, 即一个词的空间向量映射成一个投影算子, 映射关系如下:

$$\mathrm{map}(w) \longrightarrow \Pi_w, \tag{10}$$

这里的 $w \in V$, V 是字典集合, 其中 $\Pi_w = |u_w\rangle\langle u_w|$.

例如, 字典大小 $N = 3$, $V = \{$量子, 语言, 模型$\}$, 如果文档中 $d = \{$语言, 模型$\}$, 那么对应的投影算子为 $\{\Pi_2$ 和 $\Pi_3\}$, 如果单个词的空间向量用 one-hot 表示, 那么,

$$\Pi_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \Pi_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{11}$$

Π_2 和 Π_3 同时是文档 d 对应在量子概率空间中的基本量子事件.

4.1.2 词组中词与词之间的依赖关系表示

为了表示词组中两个或更多词之间的联系, 可以用不同于单个词的投影算子来表示. 例如, K 是多个词组成的词组, 即 $K = \{w_1, \ldots, w_n\}$, 对应投影算子的映射表示如下:

$$\mathrm{map}(K) \longrightarrow K_{w_1,\ldots,w_n}, \quad K_{w_1,\ldots,w_n} = |k\rangle\langle k|, \quad |k\rangle = \sum_{i=1}^{K} \alpha_i |u_i\rangle. \tag{12}$$

关于系数 $\alpha_i \in \mathbb{R}$, 为了对 $|k\rangle$ 进行单位化, α_i 的取值需要满足 $\sum_i {\alpha_i}^2 = 1$. 为了用实例说明词与词之间的依赖关系, 词表还用上面所述 V. 模拟词组 $K_{1,2,3} = \{$量子, 语言, 模型$\}$ 这 3 个词之间的依赖关系, 投影算子 $K_{1,2,3} = |K_{1,2,3}\rangle\langle K_{1,2,3}|$, 其中 $|K_{1,2,3}\rangle = \sqrt{1/5}|u_1\rangle + \sqrt{1/5}|u_2\rangle + \sqrt{3/5}|u_3\rangle$, 矢量表示如图 3 所示, 图中的概率振幅 α_1, α_2 和 α_3 就分别等于 $\sqrt{1/5}$, $\sqrt{1/5}$ 和 $\sqrt{3/5}$. $|u_1\rangle$, $|u_2\rangle$ 和 $|u_3\rangle$ 分别对应单词 “量子”、“语言” 和 “模型” 的空间向量. 投影算子 $K_{1,2,3}$ 的矩阵表示如下:

$$K_{1,2,3} = \begin{bmatrix} \frac{1}{5} & \frac{1}{5} & \frac{\sqrt{3}}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{\sqrt{3}}{5} \\ \frac{\sqrt{3}}{5} & \frac{\sqrt{3}}{5} & \frac{3}{5} \end{bmatrix}. \tag{13}$$

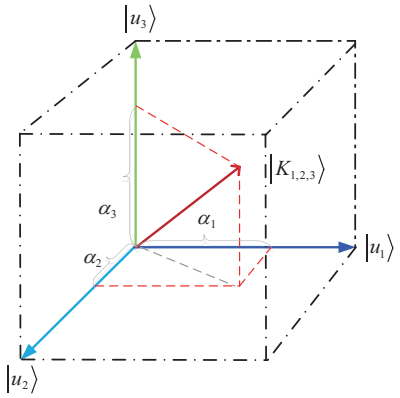


图 3 (网络版彩图) 具有依赖关系的矢量 $|K_{1,2,3}\rangle$ 的向量空间表示

Figure 3 (Color online) The dependency $|K_{1,2,3}\rangle$ is represented by vector space

4.1.3 文档表示

在传统语言模型中, 一篇文档由多个词的序列组成, 在量子语言模型中, 一篇文档用多个量子事件表示. 量子语言模型 [13] 将一篇文档 P_d 看作是 M 个量子事件 (即 M 个单词或词组), 每个量子事件用一个投影算子表示, 该文档表示为

$$P_d = \{\Pi_i : i = 1, 2, \ldots, M\}, \tag{14}$$

其中 $\Pi_i = |u\rangle\langle u|$ 表示第 i 个投影算子, $|u\rangle\langle u|$ 表示向量 $|u\rangle$ 的外积, M 是文档 d 中单词或词组的个数, $|us\rangle$ 是维度为 N 的 Hermit 空间向量, N 也是字典大小.

4.1.4 最大似然估计

对于一篇由 M 个词或词组组成的文档 $P_d = \{\Pi_1, \ldots, \Pi_M\}$, 其中 Π_i 是词或词组对应的空间向量映射得到的投影算子, 然后需要找到一种方法, 该方法能够学到一个代表这篇文档的密度矩阵 ρ. 在这项工作 [13] 中, 使用的是最大似然估计, 因为它可以自然地看成是一个经典似然函数的量子泛化. 最大似然估计可以生成一个定义明确的密度矩阵 ρ. 在量子力学中, 计算单词和词组的概率可以通过 Gleason [48] 定理计算得到, 即

$$p(\Pi_i; \rho) = \mathrm{tr}(\rho\Pi_i), \tag{15}$$

其中 $p(\Pi_i; \rho)$ 表示密度矩阵 ρ 测量各个词或词组对应的量子态 (投影算子) 的概率. 把文档中所有词或词组对应的量子态通过测量得到的概率作连乘运算, 可以得到似然函数:

$$\mathcal{L}_{P_d}(\rho) = \prod_{i=1}^{M} \mathrm{tr}(\rho\Pi_i). \tag{16}$$

最大似然函数就可以表示为

$$\mathrm{maximize}_\rho \ \log \mathcal{L}_{P_d}(\rho) = \sum_{i=1}^{M} \log \mathrm{tr}(\rho\Pi_i), \tag{17}$$

从而求得最优的密度矩阵 ρ.

4.1.5 匹配函数

Kullback Liebler (KL) 散度方法在计算不同的查询和文档表示方面具有灵活性, 这使得它对新框架中的候选评分函数很有吸引力. 经典的 KL 散度经过推广, 出现了量子相对熵 (即计算两个密度矩阵的 VN 散度). 对查询和文档分别优化出其密度矩阵 ρ_q 和 ρ_d. Sordoni 等 [13] 提出的匹配评分函数为

$$\begin{aligned} -\triangle \mathrm{VN}(\rho_q||\rho_d) &= -\mathrm{tr}(\rho_q(\log\rho_q - \log\rho_d)) \\ &\overset{\mathrm{rank}}{=} \mathrm{tr}(\rho_q \log\rho_d). \end{aligned} \tag{18}$$

实验表明在 ad-hoc 信息检索中, 相比一元语言模型和基于 MRF 的高阶语言模型, 量子语言模型取得显著的性能提升.

4.2 语音处理领域的量子语言模型

上述量子语言模型可以计算或匹配查询和文档的相关性, 但是无法建模词与词的序列关系. 不管是一段语音还是一个句子序列, 词与词之间是存在序列关系的. 语音处理领域的量子语言模型 [45] 使用了量子力学公理中的酉演化公理, 并通过连续的量子测量 (公理 2), 建模词与词的序列关系. 对于句子中所包含的词语, 不同的排列顺序可以表达不同的句子语义. 在量子语言模型中这种词语的顺序关系体现在不同的测量顺序上, 一组不同的测量顺序对应一组不同的语义表达, 文献 [45] 也是通过量子概率理论来构建语言模型, 该模型 [45] 与其他语言模型, 例如循环神经网络语言模型 [49] (recurrent neural network language model, RNNLM)、长短期记忆网络 [50] (long short-term memory, LSTM) 等相比, 在语言模型评价指标困惑度 (perplexity, PPL), 以及在自动语音识别评估设置中都取得了较好的效果.

4.2.1 量子概率的相关知识

- 一个可观测量 ρ (密度矩阵) 的投影测量输出是与可观测量的特征值 $\{\lambda_j\}$ 对应的, 即 λ_j 为密度矩阵 ρ 的特征值.
- 特征值 $\{\lambda_j\}$ 的测量输出概率为 $P(\lambda_j) = \mathrm{tr}(\rho\Pi_{\lambda_j}) = \mathrm{tr}(\Pi_{\lambda_j}\rho)$, 其中 Π_{λ_j} 是观测量 ρ 特征值 λ_j 对应的投影算子. 需要注意的是 $\Pi_{\lambda_j}^{\mathrm{T}} = \Pi_{\lambda_j}$, 同时满足 $\Pi_{\lambda_j}^2 = \Pi_{\lambda_j}$.
- 系统 ρ 对应特征值 λ_j 测量之后的状态, 用密度矩阵 ρ' 表示. 即

$$\rho' = \frac{\Pi_{\lambda_j}\rho\Pi_{\lambda_j}}{\mathrm{tr}(\Pi_{\lambda_j}\rho\Pi_{\lambda_j})},$$

其中, 分母 $\mathrm{tr}(\Pi_{\lambda_j}\rho\Pi_{\lambda_j})$ 是对密度矩阵的归一化计算.

- 演化 (实数域), 通过一个酉矩阵 U 来刻画, U 满足性质 $U^{\mathrm{T}}U = UU^{\mathrm{T}} = I$, 其中 I 是一个单位矩阵. 在 t 时刻 (第 t 次测量之后), 关于系统的演化表示为

$$\rho_{t+1} = U\rho_t U^{\mathrm{T}}.$$

关于量子概率理论的完整内容, 参见文献 [43].

4.2.2 建模过程

假设有一个长度为 n 的句子序列, $s=(w_1,w_2,\ldots,w_n)$, 模型用到的词典大小为 N, 包含模型中用到的所有单词. 然后定义 N 个标准基向量 $\{e_w : w\in\{1,2,\ldots,N\}\}$ 对应每一个词. 为了计算每一个词 w_i 出现的概率, 将每个单词对应的基向量外积运算, 作为语言模型的测量算子, 如: $\Pi_w = e_w e_w^{\mathrm{T}}$. 该模型是按次序进行测量的, 基于量子力学的测量理论, 依次测量每个单词出现的概率. 基本思想及计算步骤如算法 1. 把初始化测量得到的概率值与在循环过程中测量得到的所有条件概率

算法 1 基于量子测量的语言建模

1: **输入** 密度矩阵 ρ_0 和酉演化矩阵 U;
2: **输出** 句子序列的联合概率 $P(s|\rho_o,U)$;
3: **初始化**
投影测量概率: $P(w_1;\rho_0,U)=\mathrm{tr}(\rho_0\Pi_{w_1})$;
投影后的状态: $\rho_1'=\frac{\Pi_{w_1}\rho_0\Pi_{w_1}}{\mathrm{tr}(\Pi w_1\rho_0\Pi_{w_1})}$;
演化后的状态: $\rho_1=U\rho_1'U^{\mathrm{T}}$;
4: **循环测量** $(i=2,\ldots,n)$
投影测量概率: $P(w_i|w_1,\ldots,w_{i-1};\rho_0,U)=\mathrm{tr}(\rho_{i-1}\Pi_{w_i})$;
投影后的状态: $\rho_i'=\frac{\Pi_{w_i}\rho_{i-1}\Pi_{w_i}}{\mathrm{tr}(\Pi w_i\rho_{i-1}\Pi_{w_i})}$;
演化后的状态: $\rho_i=U\rho_i'U^{\mathrm{T}}$;
5: **结束**
$P(s|\rho_0,U)=P(w_1;\rho_0,U)\prod_{i=1}^{n}P(w_i|w_1,\ldots,w_{i-1};\rho_0,U)$;

$P(w_i|w_1,\ldots,w_{i-1};\rho_0,U)$ 连乘, 就得到序列的联合概率 $P(s|\rho_0,U)$, 其中, 密度矩阵 ρ_0 和演化酉矩阵 U 都是模型的参数, 即需要得到的模型. 最后用一个公开评价函数 PPL 评价语言模型的好坏. 评价指标函数 PPL 见下式:

$$\mathrm{PPL}(\rho_0,U)=\exp\left(-\frac{1}{|C|}\sum_{s\in C}\log P(s|\rho_0,U)\right). \tag{19}$$

C 是语料集, $|C|$ 是语料集中句子的个数, 该量子语言模型的目标是找到一组好的参数 (ρ_0 和 U) 和最小化困惑度 PPL.

4.2.3 辅助系统

在多次测量的过程中, 为了解决不丢失全局语义信息的问题, 文献 [43] 通过构建一个虚构 D 维的 Hilbert 空间 $\mathcal{H}_{\mathrm{ancilla}}=\mathcal{C}^D$, 将 $\mathcal{H}_{\mathrm{ancilla}}$ 叫做辅助系统. 然后, 将原始的空间和虚构的空间进行张量积计算, 就得到了一个 DN 维的 Hilbert 空间 $\mathcal{H}_2=\mathcal{H}_{\mathrm{ancilla}}\otimes\mathcal{H}=\mathcal{C}^{DN}$. 在这个新的空间中, 投影算子可以重新定义为 $\Pi_{w_2}^{(2)}=I_D\otimes\Pi_w$, 这里的 I_D 是一个 D 维的单位矩阵. 这种设计将耦合系统的时间演化在两个纠缠系统之间建立了非平凡的关联, 使得单词的量子态表示的测量和演化保存了一些关于整个序列的信息, 并存储在了状态表示的辅助部分中. 然后, 通过时间演变将这些信息重新转换为单个词的量子状态表示, 从而产生 “记忆效应”, 将整个句子序列考虑在内, 进而扩展了 n-gram 语言模型方法背后的思想. 但是, 这种做法也导致了大量的参数需要学习.

4.2.4 酉演化矩阵的构建

系统的演化最重要的是构建演化矩阵, 构建酉矩阵的方法有多种. 第一种方法是使用一个固定的 $(\mathrm{DN})^2$ 的实数域的酉矩阵计算 [51], 但是这种方法并不能取得一个好的 “记忆” 效果. 第二种方法是对

于每个单词都设计一个酉演化的矩阵, 很显然这种方法将导致大量的参数需要学习和优化. 受 Markov 模型 [12] 的启发, 每一个单词可以表示为一个固定维长度的词向量表示, 因此可以针对单个词对应词向量的每一维都定义一个酉矩阵. 假设一个单词的词向量表示为 $w \to (\alpha_1(w), \ldots, \alpha_p(w))$, p 表示词向量的维度. 这样就可以设置一组酉矩阵 $\boldsymbol{U} = (U_1, \ldots, U_p)$, 对于每一个单词, 我们就可以动态地构建新的酉矩阵表示:

$$V(w) = \prod_{i=1}^{p} U_i^{\alpha_i(w)}.$$

即使这样, 单词的词向量表示的向量维度一般也比较大, 因此也存在很多的参数需要学习优化.

4.3 自动问答领域的量子语言模型

上述量子语言模型存在以下不足: 第一, 在量子语言模型 [13] 模拟每个词语的时候, 使用的是独热向量 (one-hot vector) 表示的, 这种表示只能够表示该词的出现以及位置信息, 并不能有效地把全局的语义信息给建模出来; 第二, 表示单词用固定的密度矩阵表示, 该密度矩阵使用分析得到, 并不是通过训练优化得到, 它不能够通过端到端学习方法学习获得; 第三, 不能够将文本的表示、匹配, 以及排序结合起来, 而是分开进行的, 所以就不能联合优化, 从而限制了应用的推广. 为了建模全局语义表示, 匹配以及排序过程结合起来进行优化, 提高量子语言模型的表现性能, 拓宽量子语言模型的应用. 同时针对以上量子语言模型的不足, 文献 [46] 提出结合卷积神经网络 (convolution neural network, CNN) 训练的量子语言模型框架 [46] 用来做问答匹配任务. 该量子语言模型能够将需要更新的密度矩阵表示、匹配函数和训练过程统一优化, 并且在问答任务上该语言模型实验结果达到了 state-of-the-art 的语言模型接近的效果.

4.3.1 密度矩阵的句子表示

端到端的类量子语言模型 [46] 也用一个密度矩阵来表示一个句子. 与以往不同的是借助词向量 (word embeding) [52] 表示句子中的每一个单词, 这样可以建模全局的语义信息到密度矩阵中, 优于之前使用的 one-hot 表示. 假设一个句子有 N 个词, 每个词用 d 维的词向量来表示, 句子中的每个词用 $|v_i\rangle$ 来表示, 其 $|v_i\rangle \in \mathbb{R}^{d\times 1}$. 根据密度矩阵的性质, 我们可以用式 (20):

$$\rho = \sum_i p_i \Pi_i = \sum_i \lambda_i |v_i\rangle\langle v_i|, \tag{20}$$

其中 $\sum_i p_i = 1$, ρ 是一个对称矩阵.

这种表示如图 4, 该表示方法简单, 容易应用. 为了获得单位状态向量, 需要单位化词向量的分布式表示. 式 (20) 获得的密度矩阵可以表示一个句子, 这与量子系统状态相对应, 并且能够有效地表达单词与单词之间的依赖关系. 式 (20) 中的 p_i 表示一个单词的位置信息. 在实验中有比较重要的意义.

4.3.2 问句与答句的联合表示

为了做文本匹配 (问答句对的匹配) 需要计算文本之间的相似性, 基于神经网络的端到端语言模型用迹内积的公式计算问答句对的相似性, 迹内积的编码形式对应密度矩阵的联合表示, 对应问句与答句的密度矩阵表示 ρ_q 和 ρ_d. 利用式 (21) 得到问句与答句的联合表示. 问句与答句之间的联合表示如图 5.

$$M_{qa} = \rho_q \rho_a. \tag{21}$$

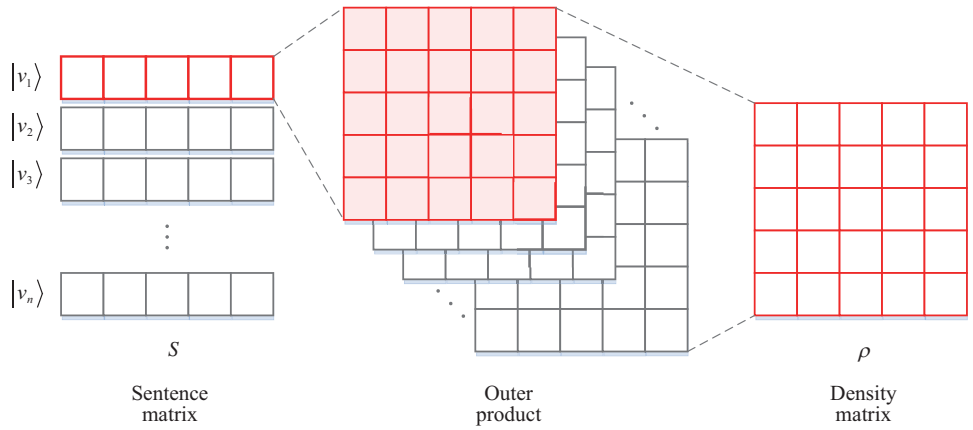


图 4 (网络版彩图) 单个句子的密度矩阵表示

Figure 4 (Color online) Single sentence representation by density matrix

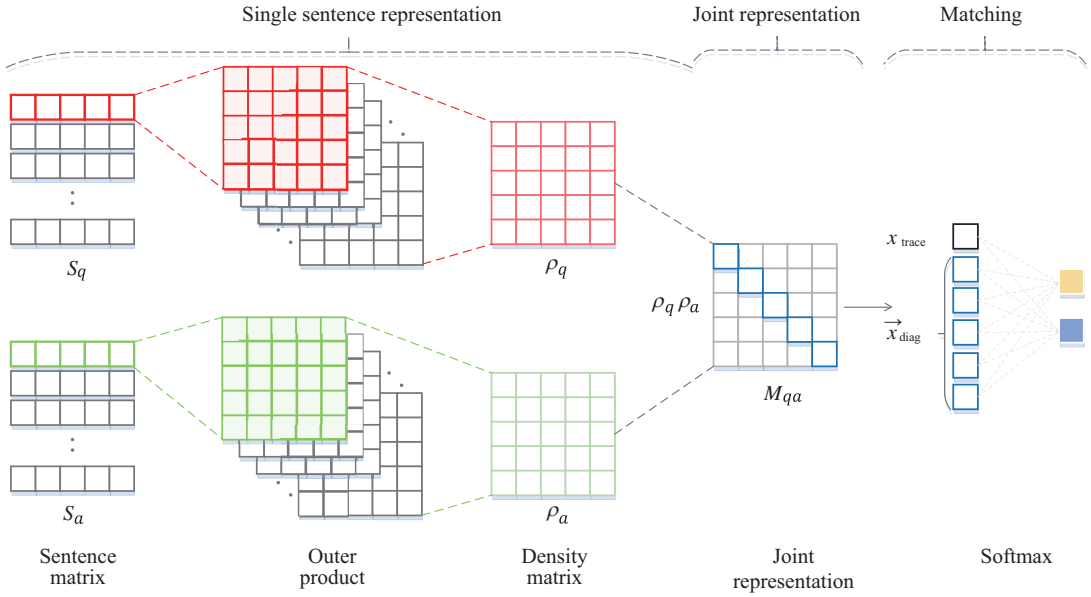


图 5 (网络版彩图) 前 3 层是句子的密度矩阵表示, 第 4 层是联合表示, 第 5 层是问答对匹配训练 softmax 输出层

Figure 5 (Color online) The first three layers are to obtain the single sentence representation, the fourth layer is to obtain the joint representation of a QA pair, and the softmax layer is to match the QA pair.

为了分析这种联合表示能够对应迹内积, 对密度矩阵做谱分解, 可以得到

$$\begin{aligned}\rho_q\rho_a &= \sum_{i,j}\lambda_i\lambda_j|v_i\rangle\langle v_j||v_j\rangle\langle v_j| \\ &= \sum_{i,j}\lambda_i\lambda_j\langle v_i|v_j\rangle|v_i\rangle\langle v_j|,\end{aligned} \tag{22}$$

其中 λ_i 是对应密度矩阵的特征值, $|v_i\rangle$ 对应特征值的特征向量. 在式 (22) 中, $\langle v_i|v_j\rangle$ 表示 $|v_i\rangle$ 和 $\langle v_j|$ 之间的相似度. 因为 $\langle v_i|v_j\rangle = \mathrm{tr}(|v_i\rangle\langle v_j|)$, 因此可以写出迹内积的公式:

$$\mathrm{tr}(\rho_q\rho_a) = \sum_{i,j} \lambda_i\lambda_j\langle v_i|v_j\rangle^2. \tag{23}$$

联合矩阵对角线元素对应问句与答句的语义重叠, 可以计算问句与答句之间的匹配相似度. 因此联合矩阵采用式 (23) 这样一种迹内积的编码方式, 计算问句与答句之间语义相似性, 然后进行问题与答案的相似度匹配. 这种表示是一个更一般化的问题与答案对的特征表示方法, 是接近 VN 散度 [13] 的表示方法.

4.3.3 学习如何匹配密度矩阵

基于神经网络的量子语言模型提出了两个方法匹配问答对. 第一种方法是迹内积的编码方式. 计算密度矩阵 ρ_q 和 ρ_a 乘积的迹, 如式 (24) 所示:

$$S(\rho_q, \rho_a) = \mathrm{tr}(\rho_q\rho_a). \tag{24}$$

这种迹内积的表示可以计算单词和句子之间相似度. 式 (24) 可以展开表示为

$$x_{\mathrm{trace}} = \mathrm{tr}(\rho_q\rho_a) = \mathrm{tr}\bigg(\sum_{i,j}\lambda_i\lambda_j|r_i\rangle\langle r_j|\bigg), \tag{25}$$

x_{trace} 可以理解为语义的叠加, 在这里表示问题和答案对应密度矩阵的相似度. 另外, 由于联合矩阵的对角线元素对应了丰富的特征信息, 对角线元素用向量 $\vec{x}_{\mathrm{diag}}$, $\vec{x}_{\mathrm{diag}}$ 中对应不同的对角线元素表示相似度测量的不同等级. 特征表示能够被定义为

$$\vec{x}_{\mathrm{feat}} = [x_{\mathrm{trace}}; \vec{x}_{\mathrm{diag}}],$$

对应神经网络的反向传播的损失函数为

$$L = -\sum_i N[y_i \log h(\vec{x}_{\mathrm{feat}}) + (1-y_i)\log 1 - h(\vec{x}_{\mathrm{feat}})],$$

其中 $h(\vec{x}_{\mathrm{feat}})$ 是 softmax 激活层之后的输出值, y_i 是问答句对的标签 (label).

另外一种方法是利用卷积神经网络的方法对联合表示做卷积操作, 这样就搭建起了一个神经网络结构, 然后做匹配训练, 这样一种结构能够达到一个很好的实验结果. 为了学习联合矩阵的抽象表示. 采用了一个二维的卷积神经网络提取联合表示的主要特征, 这些特征表示文本之间的相似距离. 这种结构的表示, 如图 6 所示. 基于神经网络的量子语言模型结合词向量的有效表示能力构建密度矩阵, 解决了密度矩阵的稀疏问题, 又结合了神经网络强有力的学习能力, 在 WIKI-QA 数据集 [53] 上显著提升了量子语言模型的效果, 并在 TREC-QA 数据集 [54] 上接近了 state-of-the-art 语言模型的效果. 但是, 上述端到端的量子语言模型, 未能说明量子力学与神经网络的本质联系, 还需要进一步的思考与研究.

5 未来愿景

鉴于神经网络建模语言模型是现在的研究热点, 量子力学的概率理论也可以建模语言模型, 那么我们可以想到, 如何构建量子力学、语言模型和神经网络三者之间的联系. 如图 7 所示的等价类图示, 等价类的说法可能为时尚早, 但是未来工作可以朝这个方向进行努力. 一些想法如下:

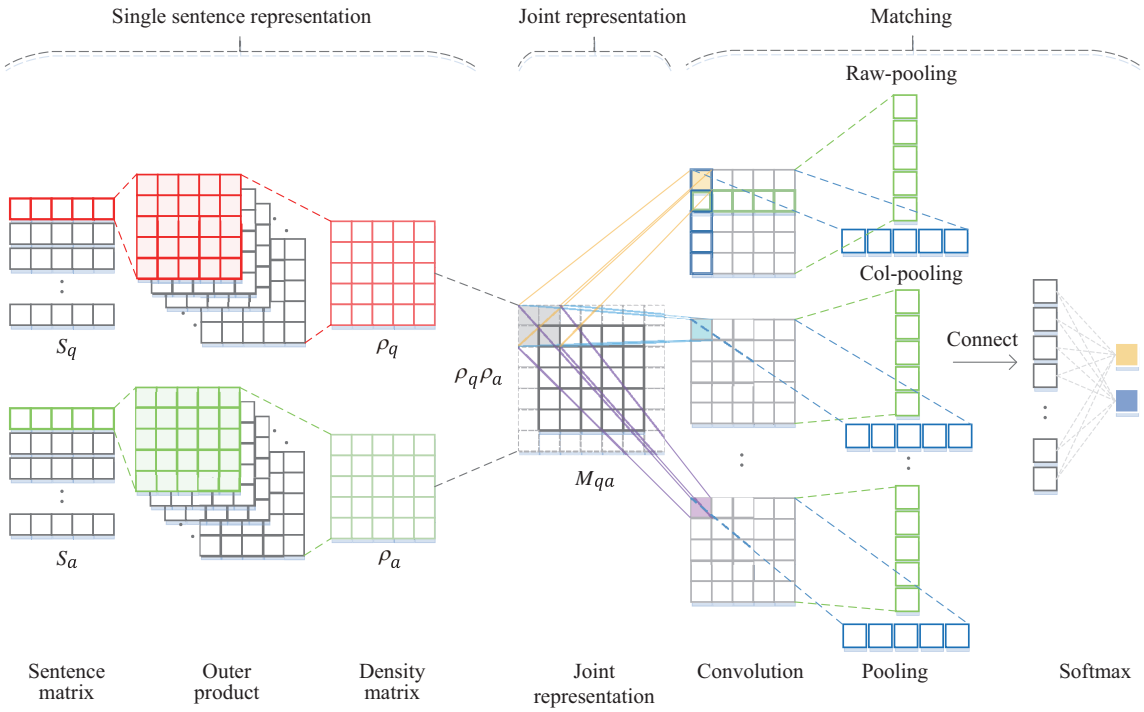


图 6　(网络版彩图) 句子表示和 QA 句对的联合表示, 以及 2 维卷积神经网络示意图

Figure 6　(Color online) The single sentence representation and the joint representation, and the rest layers are to match the QA pair by the similarity patterns learned by 2-dimensional-CNN

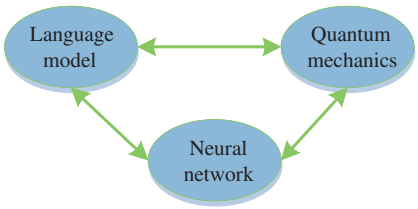


图 7　(网络版彩图) 语言建模、神经网络和量子力学的等价类图示

Figure 7　(Color online) Language modeling, neural networks and equivalent class diagrams of quantum mechanics

端到端的量子语言模型 [46] (NNQLM) 并没有解决神经网络与量子语言模型的内在联系, 因此有必要研究量子力学与神经网络之间的关系. 已有相关文献着重研究了量子力学与神经网络之间的本质关系 [55~59]. 其中发表在 *Science* 的论文 [55] 指出量子多体波函数与神经网络是有联系的, 文献 [56] 建立了量子力学领域与深度学习领域之间的基本联系, 利用这种联系来说明卷积网络的每一层信道所起的作用, 证明了深度卷积网络的实现与张量结构的量子多体波函数之间是存在等价性的, 文献 [57] 建立了张量结构的量子多体波函数与深度循环网络结构之间的联系, 这种联系可以借助张量分解构建. 文献 [58] 提出了一种基于算术电路的深层网络体系结构, 该体系结构建立了网络与分层张量分解之间的等价关系. 文献 [56~58] 都有提到量子多体波函数的高维张量结构通过张量分解可以得到一个张量网络结构. 在文献 [56,57] 启发下, 我们提出一个新的由量子多体波函数启发的语言模型, 该模型采用张量积来建模词语之间的交互, 揭示了量子语言建模中使用卷积神经网络的必要性 [59].

为了使得语言模型、神经网络以及量子力学三者在理论上建立一个如图 7 的等价类关系, 我们可以在未来的研究工作中做以下尝试. 首先我们可以用量子多体波函数进行语言建模, 每个词的波函数作张量积形成量子多体波函数表示的复合系统, 进而表示一系列的词. 量子多体波函数是用高维张量表示的, 我们需要对它进行分解, 分解之后可以得到张量网络表示. 张量网络在数学和物理领域应用广泛, 可解释性强, 所以用张量网络结构来指导神经网络建模语言, 神经网络就是一个可解释的网络. 在量子力学中, 复合系统的两个部分 (问句与答句) 的纠缠关系在自然语言处理的问答任务中是存在的. 因此可以尝试借助张量网络实现量子纠缠的语言建模, 进一步地, 基于量子语言模型的文本生成任务也将成为可能. 复数域词向量 [60] 的出现, 也启发我们在未来的工作中可以研究复数域的量子语言模型.

6 总结

本文从量子公理入手, 阐述了量子力学 4 个基本公理及其与语言建模的关系. 然后, 基于该公理体系, 详细介绍了 3 种量子语言模型, 分别是信息检索领域用于文本匹配的量子语言模型, 语音处理领域用于计算词语序列条件概率的量子语言模型, 以及在自动问答匹配任务中提出的端到端的量子语言模型. 为建立量子语言模型与神经网络之间的本质联系, 我们提出进一步的研究思路, 即利用量子多体波函数建模语言. 进而, 使得语言模型、神经网络和量子力学三者能够逐步建立等价类关系, 旨在从基础理论和实际应用等各个方面, 更好地建模语言, 并使之应用于更多的自然语言处理任务.

参考文献

1 Minsky M. Semantic Information Processing. Cambridge: MIT Press, 1968. 440–441
2 Schank R. Conceptual Information Processing. Amsterdam: Elsevier Science Inc, 1975. 5–21
3 Bendersky M, Croft W B. Modeling higher-order term dependencies in information retrieval using query hypergraphs. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, Portland, 2012. 941–950
4 Hofmann T. Probabilistic latent semantic indexing. In: Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, 1999. 50–57
5 Harris Z S. Distributional structure. Word, 1954, 10: 146–162
6 Zhai C X. Statistical language models for information retrieval. In: Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, New York, 2007. 1: 3–4
7 Brown P F, Desouza P V, Mercer R L, et al. Class-based n-gram models of natural language. Comput Linguist, 1992, 18: 467–479
8 Deerwester S, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis. J Am Soc Inf Sci, 1990, 41: 391–407
9 Xu W, Rudnicky A. Can artificial neural networks learn language models? In: Procedings of the 6th International Conference on Spoken Language Processing, 2000
10 Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model. J Mach Learn Res, 2003, 3: 1137–1155
11 Sun F, Guo J, Lan Y, et al. Sparse word embeddings using l1 regularized online learning. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016. 2915–2921
12 Metzler D, Croft W B. A Markov random field model for term dependencies. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Salvador, 2005. 472–479
13 Sordoni A, Nie J, Bengio Y. Modeling term dependencies with quantum language models for IR. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, 2013. 653–662
14 Robins D. Interactive information retrieval: context and basic notions. J Inform Sci, 2000, 3: 57–62

15 Magerman D M. Statistical decision-tree models for parsing. In: Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics, Cambridge, 1995. 276–283
16 Bahl L R, Brown P F, de Souza P V, et al. A tree-based statistical language model for natural language speech recognition. IEEE Trans Acoust Speech Signal Process, 1989, 37: 1001–1008
17 Rosenfeld R, Carbonell J G, Rudnicky A, et al. Adaptive statistical language modeling: a maximum entropy approach. Dissertation for Ph.D. Degree. Washington: Naval Research Laboratory, 2005
18 Wang J C, Xiao R, Sun Z X, et al. Research progress of web information retrieval. Comput Res Develop, 2001, 2: 187–193 [王继成, 萧嵘, 孙正兴, 等. Web 信息检索研究进展. 计算机研究与发展, 2001, 2: 187–193]
19 Manning C D, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge: Cambridge University, 2008, 151: 5
20 Salton G, Fox E A, Wu H. Extended Boolean information retrieval. Commun ACM, 1983, 26: 1022–1036
21 Salton G, Wong A, Yang C S. A vector space model for automatic indexing. Commun ACM, 1975, 18: 613–620
22 Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF. J Documentation, 2004, 60: 503–520
23 Fuhr N. Probabilistic models in information retrieval. Comput J, 1992, 35: 243–255
24 Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. J Found Trends Inf Ret, 2009, 3: 333–389
25 Lafferty J, Zhai C. Document language models, query models, and risk minimization for information retrieval. In: Proceedings of the 24th Annual International ACM SIGIR Conference, Princeton, 2001. 111–119
26 Zhai C, Lafferty J. A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proceedings of the 24th Annual International ACM SIGIR Conference, Princeton, 2001. 334–342
27 Sennrich R. Perplexity minimization for translation model domain adaptation in statistical machine translation. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, 2012. 539–549
28 Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. J Mach Learn Res, 2003, 3: 993–1022
29 Zhao Q, Tong L, Swami A, et al. Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: a POMDP framework. IEEE J Sel Areas Commun, 2007, 25: 589–600
30 van Rijsbergen C J. The Geometry of Information Retrieval. Cambridge: Cambridge University Press, 2004. 15–20
31 Zhang P, Song D W, Hou Y X, et al. Automata modeling for cognitive interference in users relevance judgment. In: Proceedings of Symposium on Quantum Informatics for Cognitive, Social, and Semantic Processes, 2010. 125–133
32 Wang B, Zhang P, Li J, et al. Exploration of quantum interference in document relevance judgement discrepancy. Entropy, 2016, 18: 144
33 Zuccon G, Azzopardi L, van Rijsbergen K. The Quantum Probability Ranking Principle for Information Retrieval. Berlin: Springer, 2009. 232–240
34 Sordoni A, He J, Nie J. Modeling latent topic interactions using quantum interference for information retrieval. In: Proceedings of the 22nd CIKM, 2013. 1197–1200
35 Zhang P, Li J, Wang B, et al. A quantum query expansion approach for session search. Entropy, 2016, 18: 146
36 Zhang P, Song D W, Zhao X Z, et al. Investigating query-drift problem from a novel perspective of Photon polarization. Berlin: Springer, 2011, 6931: 332–336
37 Zhao X, Zhang P, Song D, et al. A novel re-ranking approach inspired by quantum measurement. In: Proceedings of European Conference on Information Retrieval. Berlin: Springer, 2011. 721–724
38 Xie M J, Hou Y X, Zhang P, et al. Modeling quantum entanglements in quantum language models. In: Proceedings of the International Joint Conferences on Artificial Intelligence, 2015. 1362–1368
39 Piwowarski B, Frommholz I, Lalmas M. What can quantum theory bring to information retrieval. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 2010. 59–68
40 Frommholz I, Larsen B, Piwowarski B, et al. Supporting poly representation in a quantum-inspired geometrical retrieval framework. In: Proceedings of the 3rd Symposium on Information Interaction in Context, 2010. 115–124
41 Haven E, Khrennikov A. Quantum Social Science. Cambridge: Cambridge University Press, 2013
42 Bruza P D, Wang Z, Busemeyer J R. Quantum cognition: a new theoretical approach to psychology. Trends Cogn Sci, 2015, 19: 383–393

43 Nielsen M A, Chuang I L. Quantum Computation and Quantum Information. Cambridge: Cambridge University Press, 2000 [赵千川, 译. 量子计算和量子信息 (一). 北京: 清华大学出版社, 2004. 44–88]

44 von Neumann J. Mathematical Foundations of Quantum Mechanics. Princeton: Princeton University Press, 1996

45 Basile I, Tamburini F. Towards quantum language models. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017. 1840–1849

46 Zhang P, Niu J B, Su Z, et al. End-to-End quantum-like language models with application to question answering. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, 2018

47 Lvovsky A I. Iterative maximum-likelihood reconstruction in quantum homodyne tomography. J Opt B-Quantum Semiclass Opt, 2004, 6: S556–S559

48 van Rijsbergen C J. The Geometry of Information Retrieval. Cambridge: Cambridge University Press, 2004. 39–40

49 Shi Y Z, Zhang W Q, Cai M, et al. Variance regularization of RNNLM for speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, 2014. 4893–4897

50 Greff K, Srivastava R K, Koutník J, et al. LSTM: a search space odyssey. IEEE Trans neural Netw Learn Syst, 2017, 28: 2222–2232

51 Spengler C, Huber M, Hiesmayr B C. A composite parameterization of unitary groups, density matrices and subspaces. J Phys A-Math Theor, 2010, 43: 385306

52 Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014. 1532–1543

53 Yang Y, Yih W, Meek C. Wikiqa: a challenge dataset for open-domain question answering. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015. 2013–2018

54 Wang M, Smith N A, Mitamura T. What is the Jeopardy model? A quasi-synchronous grammar for QA. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007

55 Carleo G, Troyer M. Solving the quantum many-body problem with artificial neural networks. Science, 2017, 355: 602–606

56 Levine Y, Yakira D, Cohen N, et al. Deep learning and quantum entanglement: fundamental connections with implications to network design. In: Proceedings of the 6th International Conference on Learning Representations, 2018

57 Levine Y, Sharir O, Shashua A. Benefits of depth for long-term memory of recurrent networks. In: Proceedings of the 6th International Conference on Learning Representations, 2018

58 Cohen N, Sharir O, Shashua A. On the expressive power of deep learning: a tensor analysis. In: Proceedings of Conference on Learning Theory, 2016. 698–728

59 Zhang P, Su Z, Zhang L P, et al. A quantum many-body wave function inspired language modeling approach. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018

60 Li Q C, Uprety S, Wang B Y, et al. Quantum-inspired complex word embedding. In: Proceedings of the 3rd Workshop on Representation Learning for NLP, 2018

A survey of quantum language models

Peng ZHANG[1*], Xindian MA[1] & Dawei SONG[2]

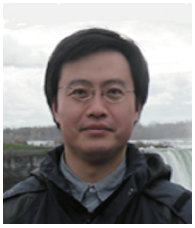
1. *School of Computer Science and Technology, Tianjin University, Tianjin* 300350*, China*;
2. *School of Computer Science & Technology, Beijing Institute of Technology, Beijing* 100081*, China*
* Corresponding author. E-mail: pzhang@tju.edu.cn

Abstract Language model is a fundamental research topic in areas related to natural language processing. In recent years, researchers have proposed quantum language models based on the probability theory of quantum mechanics. This paper aims to review the research motivation and the current progress of constructing various quantum language models. First, it reviews the research problems of classical language models. Then, it introduces some quantum language models in information retrieval and speech processing, as well as an end-to-end quantum language model based on neural network architecture. By analyzing the advantages and disadvantages of each quantum language model considered here, taking into account the essential connection between quantum mechanics and neural networks, we outline our vision for future research directions.

Keywords language model, quantum language model, neural network, information retrieval, quantum mechanics

Peng ZHANG was born in 1983. He received his Ph.D. degree from Robert Gordon University, Aberdeen, United Kingdom, in 2013. Currently, he is an associate professor in the School of Computer Science and Technology, Tianjin University. His research interests include information retrieval, natural language processing, and quantum language models.

Xindian MA was born in 1995. He received his Bachelor degree from Shanxi Agricultural University, Shanxi, China, in 2017. Currently, he is a postgraduate student at Tianjin University. His research interest focuses on quantum language models.



Dawei SONG was born in 1972. He received his Ph.D. degree from the Chinese University of Hong Kong. Currently, he is a professor in the School of Computer Science and Technology, Beijing Institute of Technology. His research interests include information retrieval, natural language processing, and quantum cognition.