End-to-End quantum language model with Application to Question Answering

Peng Zhang^[1], JiaBin Niu^[1], Zhan Su^[1], **Bengyou Wang**^[2], Liqun Ma^[1], Dawei Song^[1] Tianjin University^[1] Tencent^[2]

Contents

QA System Statistical Language Model Quantum Language Model NN-based Quantum Language Model Quantum Al

QA System



QA system in Tencent

Community QA

• FAQs

≻KBQA

Knowledge Base

➢ Passage QA

Only unstructured documents

Two-step Architecture in Community QA



Textual Matching



✓ Unsupervised Models
 ✓ TFIDF/BM25
 ✓ language model

✓ Neural Network Models
 ✓ DSSM
 ✓ CNN/RNN variants

Contents

QA System
Statistical Language Model
Quantum Language Model
NN-based Quantum Language Model
Quantum Al

Statistical Language Model

 For a sequence of terms in the document d=w1w2...wn, SLM calculates the probability P(w1w2...wn). Based on Beyes' rule, we have:

$$p(w_1 w_2 \cdots w_n) = p(w_1) p(w_2 \cdots w_n | w_1)$$
$$= p(w_1) \prod_{i=2}^n p(w_i | w_{i-1} \cdots w_1)$$

SLM-based IR model (SLMIR)

- Query likelihood model: define the relevance as the generative probability of the current query w.r.t. each document.
- Translation model: define the relevance as the probability that the query would have been generated as a translation of the document, and factor in the user's general preferences in the form of a prior distribution over documents.
- KL-divergence model: query and document are correspond to two different languages. And the relevance is defined as the KL-divergence between the two language models.
- We focus on KL-divergence model in this talk.

Contents

QA System
Statistical Language Model
Quantum Language Model
NN-based Quantum Language Model
Quantum Al

Quantum Concept

• Simple Example:



 $\vec{v} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^{\mathrm{T}}$ Projection_1 = $s_1 \cdot \vec{v} = \left(\frac{1}{\sqrt{2}}, 0\right)^{\mathrm{T}}$ Projection_2 = $s_2 \cdot \vec{v} = \left(0, \frac{1}{\sqrt{2}}\right)^{\mathrm{T}}$ Projection_3 = $s_3 \cdot \vec{v} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^{\mathrm{T}}$

- ✓ A unit vector $\vec{u} \in \mathbb{R}^n$, $\|\vec{u}\|_2 = 1$ is written as $|u\rangle$ (ket)
- ✓ The transpose \vec{u}^{T} is written as $\langle u |$ (bra)
- \checkmark The projector onto the direction \vec{u} writes as $|u\rangle\langle u|$ (dyad), corresponding to the pure state
- ✓ The inner product between two vectors writes as $\langle u|u \rangle$
- \checkmark The elements of the standard basis in \mathbb{R}^n are denoted as $|e_i\rangle = (\delta_{1i}, ..., \delta_{ni})^T$, where $\delta_{ij} = 1$, iff i = j
- ✓ Generally, any ket $|v\rangle = \sum_i v_i |u_i\rangle$ is called a superposition of the $|u_i\rangle$, where $\{|u_1\rangle, ..., |u_n\rangle\}$ form an orthonormal basis

Density Matrix

• Density Matrix

A density matrix corresponds to the discrete probability distribution in classical probability theory. It assigns a quantum probability to each one of the infinite dyads (an elementary event in quantum probability). For a density matrix ρ .

$$\rho = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

$$\mu_{\rho}(|e\rangle\langle e|) = tr(\rho|e\rangle\langle e|) = 0.5, \ \mu_{\rho}(|f\rangle\langle f|) = tr(\rho|f\rangle\langle f|) = 1$$

where: $|e\rangle = (1,0)^{\mathsf{T}}$ $|e\rangle\langle e| = \begin{bmatrix} 1 & 0\\ 0 & 0 \end{bmatrix}$ $|f\rangle = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^{\mathsf{T}}$ $|f\rangle\langle f| = \begin{bmatrix} 0.5 & 0.5\\ 0.5 & 0.5 \end{bmatrix}$

Gleason's Theorem : A. Gleason. Measures on the closed subspaces of a hilbert space. Journ. Math. Mech., 6:885–893, 1957.

Quantum Language Models (QLM)

• For example:

V = {computer, architecture, system}, *W*_d={computer, architecture}

◆If we only observe single words:

$$\mathcal{P}_{d} = \{ \mathcal{E}_{computer}, \mathcal{E}_{architecture} \}$$

$$\mathcal{E}_{computer} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \mathcal{E}_{computer} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

◆ If we observe the dependency of *"computer"* and *"architecture"*

$$k_{ca} = \sigma_c |e_c\rangle + \sigma_a |e_a\rangle, \text{Set } \sigma_c = \sqrt{2/3}, \sigma_a = \sqrt{1/3}$$
$$\mathcal{K}_{ca} = \begin{bmatrix} \frac{2}{3} & \frac{\sqrt{2}}{3} & 0\\ \frac{\sqrt{2}}{3} & \frac{2}{3} & 0\\ 0 & 0 & 0 \end{bmatrix}$$

N-gram in extended Vector Space

Word/Term Dependency	Computer	Architecture	System	Computer Architecture	Computer <i>System</i>	Architecture System	Computer System Architecture
Count	10	6	5	4	3	2	0
Frequency	0.33	0.2	0.166	0.133	0.1	0.066	0

$P(W \mid \theta_d) = [0.33, 0.2, 0.166, 0.133, 0.1, 0.066, 0]$

$$|V| = C_n^1 + C_n^2 + \dots + C_n^n = \sum_{i=0}^n C_n^i$$

The dimension of parameter in extended Vector Space : **o(n!)**

Term Dependency (N-gram) in QLM

Word/Term Dependenc y	Computer	Architecture	System	Computer Architecture	Computer <i>System</i>	Architecture System	Computer <i>System</i> Architecture
Projection	<i>e_c</i> =[1,0,0]	e _a =[0,1,0]	<i>e</i> _{<i>s</i>} =[0,0,1]	$k_{ac} = [\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 0]$	$k_{cs} = [\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}]$	$k_{as} = [0, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}]$	$k_{cas} = [\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}]$
Frequency	$\mathrm{Tr}(\rho e_c\rangle\langle e_c)$	$\operatorname{Tr}(\rho e_a\rangle\langle e_a)$	$\operatorname{Tr}(\rho e_s\rangle\langle e_s)$	${\rm Tr}(ho k_{ac} angle\langle k_{ac})$	${\rm Tr}(ho k_{cs} angle\langle k_{cs})$	$Tr(\rho k_{as})\langle k_{as})$	$Tr(\rho k_{cas})\langle k_{cas})$

$$\rho = \begin{bmatrix} \frac{2}{3} & \frac{\sqrt{2}}{3} & 0\\ \frac{\sqrt{2}}{3} & \frac{2}{3} & 0\\ 0 & 0 & 0 \end{bmatrix}$$

 $P(W|\theta_d) = [Tr(\rho|e_c)\langle e_c|), Tr(\rho|e_a)\langle e_a|), Tr(\rho|e_s)\langle e_s|), Tr(\rho|k_{ac})\langle k_{ac}|), Tr(\rho|k_{cs})\langle k_{cs}|), Tr(\rho|k_{as})\langle k_{as}|), Tr(\rho|k_{cas})\langle k_{cas}|)]$

The dimension of parameter in QLM: **o(n^2)**

Term Dependency in Quantum Entanglement

Definition 1. (QE): Let A be an n-qubit system in a state $|\phi_A\rangle$ and $\{A_1, A_2\}$ be a partition of A, where two disjoint parts A_1 and A_2 have 0 < k < n qubits and n - k qubits, respectively. A is entangled iff. there does NOT exist any tensor product decomposition of $|\phi_A\rangle$ such that $|\phi_A\rangle = |\phi_{A_1}\rangle \otimes |\phi_{A_2}\rangle$, where $|\phi_{A_1}\rangle$ and $|\phi_{A_2}\rangle$ are the states of A_1 and A_2 , respectively.

Definition 2. (UPD): A pattern $A = \{W_1, W_2, \dots, W_n\}$ forms the UPD pattern iff. the joint probability distribution over A cannot be unconditionally factorized, i.e., there does NOT exist any *m*-partition $\{A_1, A_2, \dots, A_m; m > 1\}$ of A, so that $p(a) = p(a_1) \cdot p(a_2) \cdots p(a_m)$, where $p(a_i)$, $i = 1, 2, \dots, m$, is the joint distribution over A_i .

Quantum Language Model (QLM)



- LM: a document *d* is represented by a sequence of terms
- QLM: *d* is represented by a sequence of quantum events (with dyads for a term or a dependency)

Computing probabilities

 $p(s) = \sum p(\varphi_i) |\langle s | \varphi_i \rangle|^2$ $= \sum p(\varphi_i) \left\langle \varphi_i | \widehat{M}_s^{\dagger} \widehat{M}_s | \varphi_i \right\rangle$ $= \sum p(\varphi_i) tr(\widehat{M}_s |\varphi_i\rangle \langle \varphi_i |)$ $= \operatorname{tr}\left(\widehat{M}_{s} \underbrace{p(\varphi_{i}) \sum |\varphi_{i}\rangle\langle\varphi_{i}|}_{\rho}\right)$ $= \operatorname{tr}(\widehat{M}_{s}\rho)$ $= \operatorname{tr}(\rho \widehat{M}_{s})$

Where $\boldsymbol{\rho}$ is a density matrix

Measurement in QLM

- A density matrix ρ to represent sentence
- Given the observed projectors $\mathcal{P}_d = \{\Pi_1, \dots, \Pi_M\}$ for sentence S, the quantum language model ρ is estimated through Maximum Likelihood Estimation, and the likelihood is represented as:

•
$$\mathcal{L}_{\mathcal{P}_d}(\rho) = \prod_{i=1}^M tr(\rho \Pi_i)$$

Maximum likelihood estimation for QLM

•Likelihood:

$$\mathcal{L}_{\mathcal{P}_d}(\rho) = \prod_{i=1}^M \operatorname{tr}(\rho \Pi_i).$$

• Estimation/Training of Density Matrix:

 $\begin{array}{ll} \underset{\rho}{\text{maximize}} & \log \mathcal{L}_{\mathcal{P}_d}(\rho) \\ \text{subject to} & \rho \in \mathcal{S}^n. \end{array}$

• Matching:

$$\Delta_{VN}(\rho_q \| \rho_d) = r \underline{ank}$$

 $-\operatorname{tr}(\rho_q(\log \rho_q - \log \rho_d))$ $\operatorname{tr}(\rho_q \log \rho_d),$

Limitation in QLM

If the two documents do not share any words, especially in short text
 Use embedding as a basic vector

• It is independent with the label .

• Training in a end-2-end network

Neural Network based Quantum Language Model

Contents

QA System
Statistical Language Model
Quantum Language Model
NN-based Quantum Language Model
Quantum Al

Simple version: NNQLM1

• Density matrix representation for sentences (q or a)



Architecture of NNQLM1



Using the product of the density matrixes as their joint representation, . The combined representations show the similarity of their density matrices.

Inter-sentence Similarities

• Since the density matrix is semi-positive, it

$$\rho_{q} = \sum_{i} \lambda_{i} |r_{i}\rangle \langle r_{i}|$$

$$\rho_{a} = \sum_{i,j} \lambda_{j} |r_{i}\rangle \langle r_{i}|r_{j}\rangle \langle r_{j}|$$

$$= \sum_{i,j} \lambda_{i}\lambda_{j} \langle r_{i}|r_{j}\rangle |r_{i}\rangle \langle r_{j}|$$



$$\operatorname{tr}(\boldsymbol{\rho}_{q}\boldsymbol{\rho}_{a}) = \sum_{i,j} \lambda_{i}\lambda_{j} \left\langle r_{i} | r_{j} \right\rangle^{2}$$

the similarity between ho_q and ho_a

Architecture in NNQLM2



Future work in Quantum-inspired NN

• Complex embedding

Richer input, higher performance

Interference in NN, Cross-modal fusion

• Entanglement in NN Connection and memory in NN

More works try to bridge the gap between **Quantum Concept and Deep learning** ^[1], It may open a new door to reveal the **black-box inner mechanism** of Neural Network

Contents

QA System
Statistical Language Model
Quantum Language Model
NN-based Quantum Language Model
Quantum Al

Exploration in Quantum AI

国务院印发《新一代人工智能发展规划》,明确提出 布局量子人工智能、自然语言处理、社会计算等领域



• Machine learning algorithm in Quantum computer

✓ Quantum-inspired models and ideas, but not depends on Quantum Computer

Quantum on general AI

- Solving the quantum many-body problem with artificial neural networks[J]. Science, 2017
- Deep Learning and Quantum Entanglement: Fundamental Connections with Implications to Network Design. ICLR 2018
- Deep complex Network. ICLR 2018
- Efficient representation of quantum many-body states with deep neural networks. Nature Communications.
 2017
- SchNet: A continuous-filter convolutional neural network for modeling quantum interactions, **NIPS 2017**

Quantum AI on Language

- End-to-End Quantum-like Language Models with Application to Question Answering. AAAI 2018.
- Modeling multi-query retrieval tasks using density matrix transformation. **SIGIR 2015**
- Modeling quantum entanglements in quantum language models. IJCAI 2015
- Learning Concept Embeddings for Query Expansion by Quantum Entropy Minimization. AAAI 2014
- Modeling latent topic interactions using quantum interference for information retrieval. CIKM 2013
- Modeling term dependencies with quantum language models for IR. SIGIR 2013
- Pure high-order word dependence mining via information geometry, **ICTIR 2011 best paper**.
- A novel re-ranking approach inspired by quantum measurement. ECIR 2011 best paper.