CNM: An Interpretable Complex-valued Network for Matching

Qiuchi Li [*]
University of Padua
Padua, Italy
qiuchili@dei.unipd.it

Benyou Wang [*]
University of Padua
Padua, Italy
wang@dei.unipd.it

Massimo Melucci [†]
University of Padua
Padua, Italy
melo@dei.unipd.it

Abstract

This paper seeks to model human language by the mathematical framework of quantum physics. With the well-designed mathematical formulations in quantum physics, this framework unifies different linguistic units in a single complex-valued vector space, e.g. words as particles in quantum states and sentences as mixed systems. A complex-valued network is built to implement this framework for semantic matching. With well-constrained complex-valued components, the network admits interpretations to explicit physical meanings. The proposed complex-valued network for matching (CNM)[1] achieves comparable performances to strong CNN and RNN baselines on two benchmarking question answering (QA) datasets.

1 Introduction

There is a growing concern on the interpretability of neural networks. Along with the increasing power of neural networks comes the challenge of interpreting the numerical representation of network components into human-understandable language. Lipton (2018) points out two important factors for a model to be interpretable, namely *post-hoc interpretability* and *transparency*. The former refers to explanations of why a model works after it is executed, while the latter concerns self-explainability of components through some mechanisms in the designing phase of the model.

We seek inspirations from quantum physics to build transparent and post-hoc interpretable networks for modeling human language. The emerging research field of cognition suggests that there exist quantum-like phenomena in human cognition (Aerts and Sozzo, 2014), especially language understanding (Bruza et al., 2008). Intuitively, a

sentence can be treated as a physical system with multiple words (like particles), and these words are usually polysemous (superposed) and correlated (entangled) with each other. Motivated by these existing works, we aim to investigate the following Research Question (RQ).

RQ1: Is it possible to model human language with the mathematical framework of quantum physics?

Towards this question, we build a novel quantum-theoretic framework for modeling language, in an attempt to capture the quantumness in the cognitive aspect of human language. The framework models different linguistic units as quantum states with the adoption of quantum probability (QP), which is the mathematical framework of quantum physics that models uncertainly on a uniform *Semantic Hilbert Space* (SHS).

Complex values are crucial in the mathematical framework of characterizing quantum physics. In order to preserve physical properties, the linguistic units have to be represented as complex vectors or matrices. This naturally gives rise to another research question:

RQ2: Can we benefit from the complex-valued representation of human language in a real natural language processing (NLP) scenario?

To this end, we formulate a linguistic unit as a complex-valued vector, and link its length and direction to different physical meanings: the length represents the relative weight of the word while the direction is viewed as a superposition state. The superposition state is further represented in an amplitude-phase manner, with amplitudes corresponding to the lexical meaning and phases implicitly reflecting the higher-level semantic aspects such as polarity, ambiguity or emotion.

In order to evaluate the above framework, we implement it as a complex-valued network (CNM) for semantic matching. The network is applied to the question answering task, which is the most

---

[*]Equal Contribution
[†]Corresponding Author
[1]https://github.com/wabyking/qnn.git

typical matching task that aims at selecting the best answer for a question from a pool of candidates. In order to facilitate local matching with n-grams of a sentence pair, we design a local matching scheme in CNM. Most of State-of-the-art QA models are mainly based on Convolution Neural Network (CNN), Recurrent Neural Network (RNN) and many variants thereof (Wang and Nyberg, 2015; Yang et al., 2016; Hu et al., 2014; Tan et al., 2015). However, with opaque structures of convolutional kernels and recurrent cells, these models are hard to understand for humans. We argue that our model is advantageous in terms of interpretability.

Our proposed CNM is transparent in that it is designed in alignment with quantum physics. Experiments on benchmarking QA datasets show that CNM has comparable performance to strong CNN and RNN baselines, whilst admitting post-hoc interpretations to human-understandable language. We therefore answer RQ1 by claiming that it is possible to model human language with the proposed quantum-theoretical framework in this paper. Furthermore, an ablation study shows that the complex-valued word embedding performs better than its real counterpart, which allows us to answer RQ2 by claiming that we benefit from the complex-valued representation of natural language on the QA task.

2 Background

Here we briefly introduce quantum probability and discuss a relevant work on quantum-inspired framework for QA.

2.1 Quantum Probability

Quantum probability provides a sound explanation for the phenomena and concepts of quantum mechanics, by formulating events as subspaces in a vector space with projective geometry.

2.1.1 Quantum Superposition

Quantum Superposition is one of the fundamental concepts in Quantum Physics, which describes the uncertainty of a single particle. In the micro world, a particle like a photon can be in multiple mutual-exclusive basis states simultaneously with a probability distribution. In a two-dimensional example, two basis vectors are denoted as $|0\rangle$ and $|1\rangle$[2].

---

[2]We here adopt the widely used Dirac notations in quantum probability, in which a *unit* vector $\vec{\mu}$ and its transpose $\vec{\mu}^T$ are denoted as a ket $|u\rangle$ and a bra $\langle u|$ respectively.

Superposition is implemented to model a general state which is a linear combination of basis vectors with complex-valued weights such that

$$|\phi\rangle = \alpha_0 |0\rangle + \alpha_1 |1\rangle, \qquad (1)$$

where α_0 and α_1 are complex scalars satisfying $0 \leq |\alpha_0|^2 \leq 1, 0 \leq |\alpha_1|^2 \leq 1$ and $|\alpha_0|^2 + |\alpha_1|^2 = 1$. It follows that $|\phi\rangle$ is defined over the complex field. When α_0 and α_1 are non-zero values, the state $|\phi\rangle$ is said to be a superposition of the states $|0\rangle$ and $|1\rangle$, and the scalars α_0 and α_1 denote the probability amplitudes of the superposition.

2.1.2 Measurement

The uncertainty of an ensemble system with multiple particles is encapsulated as a mixed state, represented by a positive semi-definite matrix with unitary trace called *density matrix*: $\rho = \sum_i^m |\phi_i\rangle \langle\phi_i|$, where $\{|\phi_i\rangle\}_{i=0}^m$ are pure states like Eq. 1. In order to infer the probabilistic properties of ρ in the state space, Gleason's theorem (Gleason, 1957; Hughes, 1992) is used to calculate probability to observe x through projection measurements $|x\rangle \langle x|$ that is a rank-one projector denoted as a outer product of $|x\rangle$.

$$p_x(\rho) = \langle x| \rho |x\rangle = tr(\rho |x\rangle \langle x|) \qquad (2)$$

The measured probability $p_x(\rho)$ is a non-negative real-valued scalar, since both ρ and $|x\rangle \langle x|$ are Hermitian. The unitary trace property guarantees $\sum_{x \in X} p_x(\rho) = 1$ for X being a set of orthogonal basis states.

2.2 Neural Network based Quantum-like Language Model (NNQLM)

Based on the density matrices representation for documents in information retrieval (Van Rijsbergen, 2004; Sordoni et al., 2013), Zhang et al. (2018a) built a neural network with density matrix for question answering. This Neural Network based Quantum Language Model (NNQLM) embeds a word as a unit vector and a sentence as a real-valued density matrix. The distance between a pair of density matrices is obtained by extracting features of their matrix multiplication in two ways: NNQLM-I directly takes the trace of the resulting matrix, while NNQLM-II applies convolutional structures on top of the matrix to determine whether the pair of sentences match or not.

NNQLM is limited in that it does not make proper use of the full potential of probabilistic

property of a density matrices. By treating density matrices as ordinary real vectors (NNQLM-I) or matrices (NNQLM-II), the full potential with complex-valued formulations is largely ignored. Meanwhile, adding convolutional layers on top of a density matrix is more of an empirical workaround than an implementation of a theoretical framework.

In contrast, a complex-valued matching network is built on top of a quantum-theoretical framework for natural language. In particular, an indirect way to measure the distance between two density matrices through trainable measurement operations, which makes advantage of the probabilistic properties of density matrices and also provides flexible matching score driven by training data.

3 Semantic Hilbert Space

Here we introduce the Semantic Hilbert Space $\mathcal{H}$ defined on a complex vector space $\mathcal{C}^n$, and three different linguistic units, namely sememes, words and word combinations on the space. The concept of semantic measurement is introduced at last.

Sememes. We assume $\mathcal{H}$ is spanned by the set of orthogonal basis states $\{|e_j\rangle\}_{j=1}^n$ for *sememes*, which are the minimum semantic units of word meanings in language universals (Goddard and Wierzbicka, 1994). The unit state $|e_j\rangle$ can be seen as a one-hot vector, i.e., the j-th element in $|e_j\rangle$ is one while other elements are zero, in order to obtain a set of orthogonal unit states. Semantic units with larger granularities are based on the set of sememe basis.

Words. Words are composed of sememes in superposition. Each word w is a superposition over all sememes $\{|e_j\rangle\}_{j=1}^n$, or equivalently a unit-length vector on $\mathcal{H}$:

$$|w\rangle = \sum_{j=1}^n r_j e^{i\phi_j} |e_j\rangle, \qquad (3)$$

i is the imaginary number with $i^2 = -1$. In the above expression, $\{r_j\}_{j=1}^n$ are non-negative real-valued amplitudes satisfying $\sum_{j=1}^n r_j{}^2 = 1$ and $\phi_j \in [-\pi, \pi]$ are the corresponding complex phases. In comparison to Eq. 1, $\{r_j e^{i\phi_j}\}_{j=0}^n$ are the polar form representation of the complex-valued scalars $\{\alpha_j\}_{j=0}^1$.

Word Combinations. We view a combination of words (e.g. phrase, n-gram, sentence or document) as a mixed system composed of individual words, and its representation is computed as follows:

$$\rho = \sum_j^m \frac{1}{m} |w_j\rangle \langle w_j|, \qquad (4)$$

where m is the number of words and $|w_j\rangle$ is word superposition state in Eq. 3, allowing multiple occurrences. Eq. 4 produces a density matrix ρ for semantic composition of words. It also describes a non-classical distribution over the set of sememes: the complex-valued off-diagonal elements describes the correlations between sememes, while the diagonal entries (guaranteed to be real by its original property) correspond to a standard probability distribution. The off-diagonal elements provide our framework some potentials to model the possible interactions between the basic sememe basis, which was usually considered mutually independent with each other.

Semantic Measurements. The high-level features of a sequence of words are extracted through measurements on its mixed state. Given a density matrix ρ of a mixed state, a rank-one projector P, which is the outer product of a unit complex vector, i.e. $P = |x\rangle \langle x|$, is applied as a measurement projector. It is worth mentioning that $|x\rangle$ could be any pure state in this Hilbert space (not only limited to a specific word w). The measured probability is computed by Gleason's Theorem in Eq. 2.

4 Complex-valued Network for Matching

We implemented an end-to-end network for matching on the Semantic Hilbert Space. Fig. 1 shows the overall structure of the proposed Complex-valued Network for Matching (CNM). Each component of the network is further discussed in this section.

4.1 Complex-valued Embedding

On the Semantic Hilbert Space, each word w is embedded as a complex-valued vector $\vec{w}$. Here we link its length and direction to different physical meanings: the length of a vector represents the relative weight of the word while the vector direction is viewed as a superposition state. Each word w adopts a normalization into a superposition state $|w\rangle$ and a word-dependent weight $\pi(w)$:

$$|w\rangle = \frac{\vec{w}}{||\vec{w}||}, \quad \pi(w) = ||\vec{w}||, \qquad (5)$$

where $||\vec{w}||$ denotes the 2-norm length of $\vec{w}$. $\pi(w)$ is used to compute the relative weight of a word in
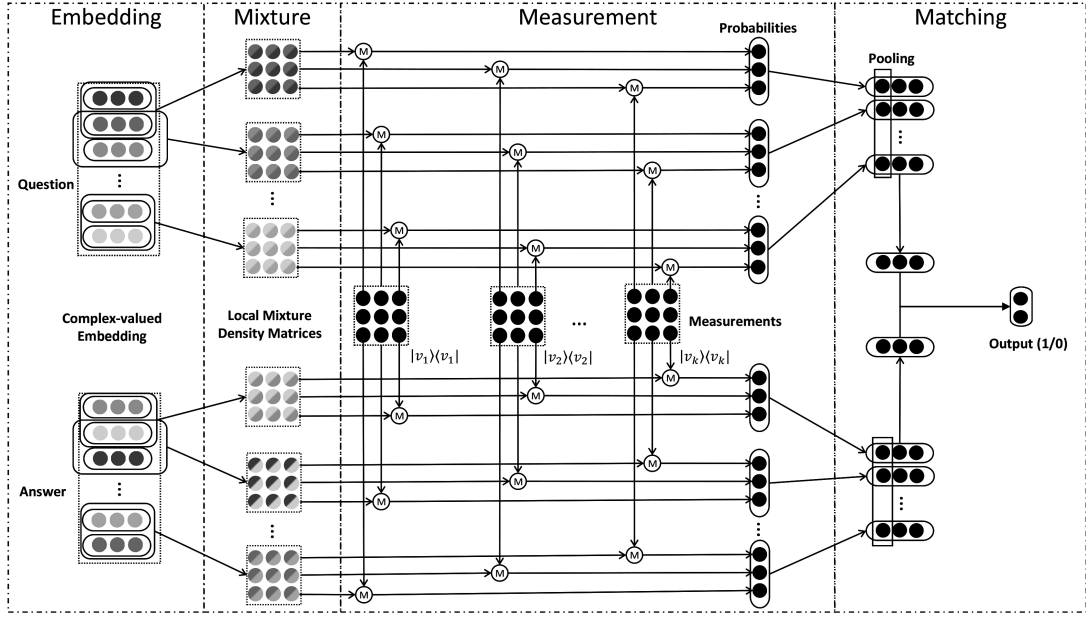
Figure 1: Architecture of Complex-valued Network for Matching. Ⓜ refers to the measurement operation in Eq. 2.

a local context window, which we will elaborate in Section 4.2.

4.2 Sentence Modeling with Local Mixture Scheme

A sentence is modeled as a combination of individual words in it. NNQLM (Zhang et al., 2018a) models a sentence as a *global mixture* of all words, which implicitly assumes a global interaction among all sentence words. This seems to be unreasonable in practice, especially for a long text segment such as a paragraph or a document, where the interaction between the first word and the last word is often negligible. Therefore, we address this limitation by proposing a *local mixture* of words, which tends to capture the semantic relations between neighboring words and undermine the long-range word dependencies. As is shown in Fig. 2, a sliding window is applied and a density matrix is constructed for a local window of length l (e.g. 3). Therefore, a sentence is composed of a sequence of density matrices for l-grams.

The representation of a local l-gram window is obtained by an improved approach over Eq. 4. In Eq. 4, each word is assigned with the same weight, which does not hold from an empirical point of view. In this study, we take the $L2$-norm of the word vector as the relative weight in a local context window for a specific word, which could be updated during training. To some extent, $L2$-norm is a measure of semantic richness of a word, i.e. the longer the vector the richer the meaning. The
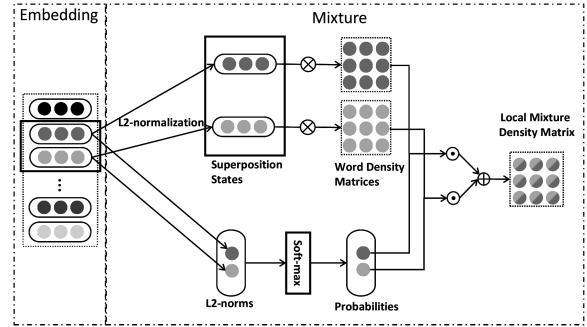

Figure 2: Architecture of local mixture component. A sliding window in the black color is applied to the sentence, generating a local mixture density matrix for each local window of length l. ⊙ means that a matrix multiplies a number with each element. ⊗ denotes the outer product of a vector with itself.

density matrix of an l-gram is computed as follows:

$$\rho = \sum_i^l p(w_i) |w_i\rangle \langle w_i|, \qquad (6)$$

where the relative importance of each word $p(w_i)$ in an l-gram is the soft-max normalized word-dependent weight: $p(w_i) = \frac{e^{\pi(w_i)}}{\sum_j^l e^{\pi(w_j)}}$, where $\pi(w_i)$ is the word-dependent weight. By converting word-dependent weights to a probability distribution, a legal density matrix is produced, because $\sum_i^l p(w_i) = 1$ gives $tr(\rho) = 1$. Moreover, the weight of a word also depends on its neighboring words in a local context.

4.3 Matching of Question and Answer

In quantum information, there have been works trying to estimate a quantum state from the results of a series of measurements (Řeháček et al., 2001; Lvovsky, 2004). Inspired by these works, we introduce trainable measurements to extract density matrix features and match a pair of sentences.

Suppose a pair of sentences with length L are represented as two sets of density matrices $\{\rho_{1j}\}_{j=1}^L$ and $\{\rho_{2j}\}_{j=1}^L$ respectively. The same set of K semantic measurement operators $\{|v_k\rangle\}_{k=1}^K$ are applied to both sets, producing a pair of k-by-L probability matrix p^1 and p^2, where $p_{jk}^1 = \langle v_k|\rho_{1j}|v_k\rangle$ and $p_{jk}^2 = \langle v_k|\rho_{2j}|v_k\rangle$ for $k \in \{1,...,K\}$ and $j \in \{1,...,L\}$. A classical vector-based distances between p^1 and p^2 can be computed as the matching score of the sentence pair. By involving a set of semantic measurements, the properties of density matrix are taken into consideration in computing the density matrix distance.

We believe that this way of computing density matrix distance is both theoretically sound and applicable in practice. The trace inner product of density matrices (Zhang et al., 2018a) breaks the basic axioms of metric, namely the non-negativity, identity of indiscernibles and triangle inequality. The CNN-based feature extraction (Zhang et al., 2018a) for density matrix multiplication loses the property of density matrix as a probability distribution. Nielsen and Chuang (2010) introduced three measures namely trace distance, fidelity, and VN-divergence. However, it is computationally costly to compute these metrics and propagate the loss in an end-to-end training framework.

We set the measurements to be trainable so that the matching of question and answering can be integrated into the whole neural network, and identify the discriminative semantic measurements in a data-driven manner. From the perspective of linear discriminant analysis (LDA) (Fisher, 1936), this approach is intended to find a group of finite discriminative projection directions for a better division of different classes, but in a more sound framework inspired by quantum probability with complex-valued values. From an empirical point of view, the data-driven measurements make it flexible to match two sentences.

Dataset	train	dev	test
TREC QA	1229/53417	65/117	68/1442
WikiQA	873/8627	126/130	633/2351

Table 1: Dataset Statistics. For each cell, the values denote the number of questions and question-answer pairs respectively.

5 Experiments

5.1 Datasets and Evaluation Metrics

The experiments were conducted on two benchmarking question answering datasets for question answering (QA), namely TREC QA (Voorhees and Tice, 2000) and WikiQA (Yang et al., 2015). TREC QA is a standard QA dataset in the Text REtrieval Conference (TREC). WikiQA is released by Microsoft Research on open domain question answering. On both datasets, the task is to select the most appropriate answer from the candidate answers for a question, which requires a ranking of candidate answers. After removing the questions with no correct answers, the statistics of the cleaned datasets are given in the Tab. 1. Two common rank-based metrics, namely mean average precision (MAP) and mean reciprocal rank (MRR), are used to measure the performance of models.

5.2 Experiment Details

5.2.1 Baselines

We conduct a comprehensive comparison across a wide range of models. On TREC QA the experimented models include Bigram-CNN (Yu et al., 2014), three-layered Long Short-term Memory (LSTM) in combination with BM25 (LSTM-3L-BM25) (Wang and Nyberg, 2015), attention-based neural matching model (aNMM) (Yang et al., 2016), Multi-perspective CNN (MP-CNN) (He et al., 2015), CNTN (Qiu and Huang, 2015), attention-based LSTM+CNN model (LSTM-CNN-attn) (Tang et al., 2015) and pairwise word interaction modeling (PWIM) (He and Lin, 2016). On WikiQA dataset, we involve the following models into comparison: Bigram-CNN (Yu et al., 2014), CNN with word count information (CNN-Cnt) (Yang et al., 2015), QA-BILSTM (Santos et al., 2016), BILSTM with attentive pooling (AP-BILSTM) (Santos et al., 2016), and LSTM with attention (LSTM-attn) (Miao et al., 2015). On both datasets, we report the results of quantum language model (Sordoni et al., 2013) and two models NNQLM-I, NNQLM-II by (Zhang et al.,

) for comparison.

5.2.2 Parameter Settings

The parameters in the network are $\Theta = \{R, \Phi, \{|v_i\rangle\}_{i=1}^k\}$, in which R and Φ denote the lookup tables for amplitudes and complex phases of each word, and $\{|v_i\rangle\}_{i=1}^k$ denotes the set of semantic measurements. We use 50-dimension complex word embedding. The amplitudes are initialized with 50-dimension Glove vectors (Pennington et al., 2014) and L2-norm regularized during training. The phases are randomly initialized under a normal distribution of $[-\pi, \pi]$. The semantic measurements $\{|v_i\rangle\}_{i=1}^k$ are initialized with orthogonal real-valued one-hot vectors, and each measurement is constrained to be of unit length during training. We perform max pooling over the sentence dimension on the measurement probability matrices, resulting in a k-dim vector for both a question and an answer. We concatenate the vectors for $l = 1, 2, 3, 4$ for questions and answers, and the larger size of windows are also tried. We will use a longer sliding window in datasets with longer sentences. The cosine similarity is used as the distance metric of measured probabilities. We use triplet hinge loss and set the margin $\alpha = 0.1$. A dropout layer is built over the embedding layer and measurement probabilities with a dropout rate of 0.9.

A grid search is conducted over the parameter pools to explore the best parameters. The parameters under exploration include $\{0.01, 0.05, 0.1\}$ for the learning rate, $\{1e - 5, 1e - 6, 1e - 7, 1e - 8\}$ for the L2-normalization of complex word embeddings, $\{8, 16, 32\}$ for batch size, and $\{50, 100, 300, 500\}$ for the number of semantic measurements.

5.2.3 Parameter Scale

The proposed CNM has a limited scale of parameters. Apart from the complex word embeddings which are $|V| \times 2n$ by size, the only set of parameters are $\{|v_i\rangle\}_{i=1}^k$ which is $k \times 2n$, with $|V|, k, n$ being the vocabulary size, number of semantic measurements and the embedding dimension, respectively. In comparison, a single-layered CNN has at least $l \times k \times n$ additional parameters with l being the filter width, while a single-layered LSTM is $4 \times k \times (k + n)$ by the minimum parameter scale. Although we use both amplitude part and phase part for word embedding, lower dimensions of embedding are adopted, namely 50, with

Model	MAP	MRR
Bigram-CNN	0.5476	0.6437
LSTM-3L-BM25	0.7134	0.7913
LSTM-CNN-attn	0.7279	0.8322
aNMM	0.7495	0.8109
MP-CNN	**0.7770**	0.8360
CNTN	0.7278	0.7831
PWIM	0.7588	0.8219
QLM	0.6780	0.7260
NNQLM-I	0.6791	0.7529
NNQLM-II	0.7589	0.8254
CNM	0.7701	**0.8591**
Over NNQLM-II	1.48% ↑	4.08% ↑

Table 2: Experiment Results on TREC QA Dataset. The best performed values are in bold.

Model	MAP	MRR
Bigram-CNN	0.6190	0.6281
QA-BILSTM	0.6557	0.6695
AP-BILSTM	0.6705	0.6842
LSTM-attn	0.6639	0.6828
CNN-Cnt	0.6520	0.6652
QLM	0.5120	0.5150
NNQLM-I	0.5462	0.5574
NNQLM-II	0.6496	0.6594
CNM	**0.6748**	**0.6864**
Over NNQLM-II	3.88% ↑	4.09% ↑

Table 3: Experiment Results on WikiQA Dataset. The best performed values for each dataset are in bold.

the comparable performance.

Therefore, our network scales better than the advanced models on the CNN or LSTM basis.

5.3 Experiment Results

Tab. 2 and 3 show the experiment results on TREC QA and WikiQA respectively, where bold values are the best performances out of all models. Our model achieves 3 best performances out of the 4 metrics on TREC QA and WikiQA, and performs slightly worse than the best-performed models on the remaining metric. This illustrates the effectiveness of our proposed model from a general perspective.

Specifically, CNM outperforms most CNN and LSTM-based models, which have more complicated structures and a relatively larger parameters scale. Also, CNM performs better than existing quantum-inspired QA models, QLM and NNQLM on both datasets, which means that the quantum theoretical framework gives rise to better performs model. Moreover, a significant improvement over NNQLM-1 is observed on these two datasets, supporting our claim that the trace inner product is not an effective distance metric of two density matrices.

4144

Setting	MAP	MRR
FastText-MaxPool	0.6659 (0.1042↓)	0.7152 (0.1439↓)
CNM-Real	0.7112 (0.0589↓)	0.7922 (0.0659↓)
CNM-Global-Mixture	0.6968 (0.0733↓)	0.7829 (0.0762↓)
CNM-trace-inner-product	0.6952 (0.0749↓)	0.7688 (0.0903↓)
CNM	0.7701	0.8591

Table 4: Ablation Test. The values in parenthesis are the performance differences between the model and CNM.

5.4 Ablation Test

An ablation test is conducted to examine the influence of each component on our proposed CNM. The following models are implemented in the ablation test. *FastText-MaxPool* adopt max pooling over word-embedding, just like FastText (Joulin et al., 2016). *CNM-Real* replaces word embeddings and measurements with their real counterparts. *CNM-Global-Mixture* adopts a global mixture of the whole sentence, in which a sentence is represented as a single density matrix, leading to a probability vector for the measurement result. *CNM-trace-inner-product* replaces the trainable measurements with trace inner product like NNQLM.

For the real-valued models, we replace the embedding with double size of dimension, in order to eliminate the impact of the parameter scale on the performance. Due to limited space, we only report the ablation test result on TREC QA, and WikiQA has similar trends. The test results in Tab. 4 demonstrate that each component plays a crucial role in the CNM model. In particular, the comparison with CNM-Real and FastText-MaxPool shows the effectiveness of introducing complex-valued components, the increase in performance over CNM-Global-Mixture reveals the superiority of local mixture, and the comparison with CNM-trace-inner-product confirms the usefulness of trainable measurements.

6 Discussions

This section aims to investigate the proposed research questions mentioned in Sec 1. For RQ1, we explain the physical meaning of each component in term of the transparency (Sec. 6.1), and design some case studies for the post-hoc interpretability (Sec. 6.2). For RQ2, we argue that the complex-valued representation can model different aspects of semantics and naturally address the non-linear semantic compositionality, as discussed in Sec. 6.3.

Components	DNN	CNM
Sememe	-	basis one-hot vector / **basis state** $\{e\|e \in \mathcal{R}^n, \|\|e\|\|_2 = 1\}$ complete &orthogonal
Word	real vector $(-\infty, \infty)$	unit complex vector / **superposition state** $\{w\|w \in \mathcal{C}^n, \|\|w\|\|_2 = 1\}$
N-gram/ Word combinations	real vector $(-\infty, \infty)$	density matrix / **mixed system** $\{\rho\|\rho = \rho^*, tr(\rho) = 1\}$
Abstraction	CNN/RNN $(-\infty, \infty)$	projector / **measurement** $\{vv^T\|v \in \mathcal{C}^n, \|\|v\|\|_2 = 1\}$
Sentence representation	real vector $(-\infty, \infty)$	real value/ **measured probability** $(0, 1)$

Table 5: Physical meanings and constraints.

	Selected words
Important	studio, president, women, philosophy scandinavian, washingtonian, berliner, championship defiance, reporting, adjusted, jarred
Unimportant	71.2, 5.5, 4m, 296036, 3.5 may, be, all, born movements, economists, revenues, computers

Table 6: Selected learned important words in TREC QA. All words are converted to lower cases.

6.1 Transparency

CNM aims to unify many semantic units with different granularity e.g. sememes, words, phrases (or N-gram) and document in a single complex-valued vector space, as shown in Tab. 5. In particular, we formulate atomic sememes as a group of complete orthogonal basis states and words as superposition states over them. A linguistic unit with larger-granularity e.g. a word phrase or a sentence is represented as a mixed system over the words (with a density matrix, i.e. a positive semi-definite matrix with unit trace).

More importantly, trainable projection measurements are used to extract high-level representation for a word phrase or a sentence. Each measurement is also directly embedded in this unified Hilbert space, as a specific unit state (like words), thus making it easily understood by the neighbor words near this specific state. The corresponding trainable components in state-of-art neural network architectures, namely, kernels in CNN and cells in RNN, are represented as arbitrary real-valued without any constraints, lead to difficulty to be understood.

6.2 Post-hoc Interpretability

The post-hoc Interpretability is shown in three groups of case studies, namely word weight scheme, matching pattern, and discriminative semantic measurements.

6.2.1 Word Weighting Scheme

Tab. 6 shows the words selected from the top-50 most important words as well as top-50 unimpor-

Question	Correct Answer
Who is the [president or chief executive of Amtrak] ?	" Long-term success ... " said George Warrington , [Amtrak 's president and chief executive] ."
When [was Florence Nightingale born] ?	," On May 12 , 1820 , the founder of modern nursing , [Florence Nightingale , was born] in Florence , Italy ."
When [was the IFC established] ?	[IFC was established in] 1956 as a member of the World Bank Group .
[how did women 's role change during the war]	..., the [World Wars started a new era for women 's] opportunities to ....
[Why did the Heaven 's Gate members commit suicide] ?,	This is not just a case of [members of the Heaven 's Gate cult committing suicide] to ...

Table 7: The matching patterns for specific sentence pairs in TREC QA. The darker the color, the bigger the word weight is. ⌈ and ⌉ denotes the possible border of the current sliding windows.

tant ones. The importance of words is based on the L2-norm of its learned amplitude embedding according to Eq. 5. It is consistent with the intuition that, the important words are more about specific topics or discriminative nouns, while the unimportant words include meaningless numbers or super-high frequency words. Note that some special form (e.g. plural form in the last row) of words are also identified as unimportant words, since we commonly did not stem the words.

6.2.2 Matching Pattern

Tab. 7 shows the match schema with local sliding windows. In a local context window, we visualize the relative weights (i.e. the weights after normalized by *softmax*) for each word with darkness degrees. The table illustrates that our model is capable of identifying true matched local windows of a sentence pair. Even some words are replaced with similar forms (e.g. commit and committing in the last case) or meanings (e.g. change and new in the fourth case), it could be robust to get a relatively high matching score. From an empirical point of view, our model outperforms other models in situations where specific matching pattern are crucial to the sentence meaning, such as when two sentences share some unordered bag-of-word combinations. To some extent, it is robust up to replacement of words with similar ones in the Semantic Hilbert Space.

6.2.3 Discriminative Semantic Measurements

The semantic measurements are performed through rank-one projectors $\{|x\rangle \langle x|\}$. From a classical point of view, each projector is associated with a superposition of fundamental sememes, which is not necessarily linked to a particular word. Since the similarity metric in the Semantic Hilbert Space can be used to indicate semantic relatedness, we rely on the nearby words of the learned measurement projectors to understand what they may refer to.

Essentially, we identified the 10 most similar words to a measurement based on the cosine sim-

	Selected neighborhood words for a measurement vector
1	andes, nagoya, inter-american, low-caste
2	cools, injection, boiling,adrift
3	andrews, paul, manson, bair
4	historically, 19th-century, genetic, hatchback
5	missile, exile, rebellion, darkness

Table 8: Selected learned measurements for TREC QA. They were selected according to nearest words for a measurement vector in Semantic Hilbert Space.

ilarity metric. Tab. 8 shows part of the most similar words of 5 measurements, which are randomly chosen from the total number of k=10 trainable measurements for the TREC QA dataset. It can be seen that the first three selected measurements were about positions, movement verbs, and people's names, while the rest were about the topic of history and rebellion respectively. Even though a clear explanation of the measurements is not available, we are still able to roughly understand the meaning of the measurements in the proposed data-driven approach.

6.3 Complex-valued Representation

In CNM, each word is naturally embedded as a complex vector, composed of a complex phase part, a unit amplitude part, and a scalar-valued length. We argue that the amplitude part (i.e. the squared root of a probabilistic weight), corresponds to the classical word embedding with the lexical meaning, while the phase part implicitly reflects the higher-level semantic aspect e.g. polarity, ambiguity or emotion. The scalar-valued length is considered as the relative weight in a mixed system. The ablation study in Sec. 5.4 confirms that the complex-valued word embedding performs better than the real word embedding, which indicates that we benefit from the complex-valued embedding on the QA task.

From a mathematical point of view, complex-valued word embedding and other complex-valued components forms a new Hilbert vector space for modelling language, with a new definitions of addition and multiplication, as well as a new inner product operation. For instance, addition in the

word meaning combination is defined as

$$
\begin{aligned}
z = & z_1 + z_2 = r_1 e^{i\theta_1} + r_2 e^{i\theta_2} \\
= & \sqrt{r_1^2 + r_2^2 + 2 r_1 r_2 \cos(\theta_2 - \theta_1)} \quad (7) \\
& \times e^{\,i \arctan\left(\frac{r_1 \sin(\theta_1) + r_2 \sin(\theta_2)}{r_1 \cos(\theta_1) + r_2 \cos(\theta_2)}\right)}
\end{aligned}
$$

where z_1 and z_2 are the values for the corresponding element for two different word vectors $|w_1\rangle$ and $|w_2\rangle$ respectively. Both the amplitudes and complex phases of z are added with a nonlinear combination of phases and amplitudes of z_1 and z_2. A classical linear addition gives $\hat{z} = r_1 + r_2$, which can be viewed as a degenerating case of the complex-valued addition with the phase information being removed ($\theta_1 = \theta_2 = 0$ in the example).

7 Conclusions and Future Work

Towards the interpretable matching issue, we propose two research questions to investigate the possibility of language modelling with quantum mathematical framework. To this end, we design a new framework to model all the linguistic units in a unified Hilbert space with well-defined mathematical constraints and explicit physical meaning. We implement the above framework with neural network and then demonstrate its effectiveness in question answering (QA) task. Due to the well-designed components, our model is advantageous with its interpretability in term of transparency and post-hoc interpretability, and also shows its potential to use complex-valued components in NLP.

Despite the effectiveness of the current network, we would like to further explore the phase part in complex-valued word embedding to directly link to concrete semantics such as word sentiment or word position. Another possible direction is to borrow other quantum concepts to capture the interaction and non-interaction between word semantics, such as the *Fock Space* (Sozzo, 2014) which considers both interacting and non-interacting entities in different Hilbert Spaces. Furthermore, a deeper and robust quantum-inspired neural architecture in a higher-dimension Hilbert space like (Zhang et al., 2018b) is also worth to be investigated for achieving stronger performances with better explanatory power.

References

Diederik Aerts and Sandro Sozzo. 2014. Quantum Entanglement in Concept Combinations. *International Journal of Theoretical Physics*, 53(10):3587–3603.

Peter D Bruza, Kirsty Kitto, Douglas McEvoy, and Cathy McEvoy. 2008. Entangling words and meaning. pages QI, 118–124. College Publications.

Ronald A Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Andrew M Gleason. 1957. Measures on the closed subspaces of a hilbert space. *J. of Math. and Mech.*, pages 885–893.

Cliff Goddard and Anna Wierzbicka. 1994. *Semantic and lexical universals: Theory and empirical findings*. John Benjamins Publishing.

Hua He, Kevin Gimpel, and Jimmy Lin. 2015. Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks. In *EMNLP*, pages 1576–1586. ACL.

Hua He and Jimmy Lin. 2016. Pairwise Word Interaction Modeling with Deep Neural Networks for Semantic Similarity Measurement. In *NAACL*, pages 937–948. ACL.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional Neural Network Architectures for Matching Natural Language Sentences. In *Advances in Neural Information Processing Systems 27*, pages 2042–2050. Curran Associates, Inc.

Richard IG Hughes. 1992. *The structure and interpretation of quantum mechanics*. Harvard university press.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Queue*, 16(3):30:31–30:57.

A I Lvovsky. 2004. Iterative maximum-likelihood reconstruction in quantum homodyne tomography. *Journal of Optics B: Quantum and Semiclassical Optics*, 6(6):S556.

Yishu Miao, Lei Yu, and Phil Blunsom. 2015. Neural Variational Inference for Text Processing. *arXiv:1511.06038*.

Michael A. Nielsen and Isaac L. Chuang. 2010. *Quantum computation and quantum information*. Cambridge University Press, Cambridge ; New York.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Xipeng Qiu and Xuanjing Huang. 2015. Convolutional neural tensor network architecture for community-based question answering. In *IJCAI*, pages 1305–1311.

Cicero dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. 2016. Attentive pooling networks. *arXiv preprint arXiv:1602.03609*.

Alessandro Sordoni, Jian-Yun Nie, and Yoshua Bengio. 2013. Modeling term dependencies with quantum language models for ir. In *SIGIR*, pages 653–662. ACM.

Sandro Sozzo. 2014. A quantum probability explanation in Fock space for borderline contradictions. *Journal of Mathematical Psychology*, 58:1–12.

Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. LSTM-based Deep Learning Models for Non-factoid Answer Selection. ArXiv: 1511.04108.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In *EMNLP*, pages 1422–1432, Lisbon, Portugal.

Cornelis Joost Van Rijsbergen. 2004. *The geometry of information retrieval*. Cambridge University Press.

J. Řeháček, Z. Hradil, and M. Ježek. 2001. Iterative algorithm for reconstruction of entangled states. *Phys. Rev. A*, 63:040303.

Ellen M Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. *SIGIR*, pages 200–207.

Di Wang and Eric Nyberg. 2015. A Long Short-Term Memory Model for Answer Sentence Selection in Question Answering. In *ACL*, pages 707–712.

Liu Yang, Qingyao Ai, Jiafeng Guo, and W. Bruce Croft. 2016. aNMM: Ranking Short Answer Texts with Attention-Based Neural Matching Model. In *CIKM*, pages 287–296. ACM.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *EMNLP*, pages 2013–2018. Association for Computational Linguistics.

Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep Learning for Answer Sentence Selection. ArXiv: 1412.1632.

Peng Zhang, Jiabin Niu, Zhan Su, Benyou Wang, Liqun Ma, and Dawei Song. 2018a. End-to-End Quantum-like Language Models with Application to Question Answering. *AAAI.*, pages 5666–5673.

Peng Zhang, Zhan Su, Lipeng Zhang, Benyou Wang, and Dawei Song. 2018b. A quantum many-body wave function inspired language modeling approach. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1303–1312. ACM.