Semantic Hilbert Space for Text Representation Learning

Benyou Wang, Qiuchi Li, Massimo Melucci
University of Padua
Padua, Italy
wang,qiuchili,melo@dei.unipd.it

Dawei Song
Beijing Institute of Technology
Beijing, China
dawei.song@open.ac.uk

ABSTRACT

Capturing the meaning of sentences has long been a challenging task. Current models tend to apply linear combinations of word features to conduct semantic composition for a bigger-granularity units e.g. phrase, sentence and documents. However, the semantic linearity does not always hold in human language. For instance, the meaning of the phrase "ivory tower" can not be deduced by linearly combining the meanings of "ivory" and "tower". To address this issue, we propose a new framework that models different levels of semantic units (e.g. sememe, word, sentence and semantic abstraction) on a single *Semantic Hilbert Space*, which naturally admits a non-linear semantic composition by means of a complex-valued vector word representation. An end-to-end neural network (https://github.com/wabyking/qnn) is proposed to implement the framework in the text classification task, and evaluation results on six benchmarking text classification datasets demonstrate the effectiveness, robustness and self-explanation power of the proposed model. Furthermore, intuitive case studies are conducted to help end users to understand how the framework works.

CCS CONCEPTS

• **Information systems** → *Document structure*; *Content analysis and feature selection*;

KEYWORDS

text understanding, neural network, quantum theory

ACM Reference Format:
Benyou Wang, Qiuchi Li, Massimo Melucci and Dawei Song. 1997. Semantic Hilbert Space for Text Representation Learning. In *Proceedings of ACM Woodstock conference (WWW '2019)*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

† Benyou Wang and Qiuchi Li contribute equally and share the co-first authorship.

WWW '2019, May 2019, San Francisco, California USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In natural language understanding, it is crucial, yet challenging, to model sentences and capture their meanings. Essentially, most statistical machine learning models [5, 16, 22, 27, 35] are built within a linear bottom-up framework, where words are the basic features adopting a low-dimensional vector representation, and a sentence is modeled as a linear combination of individual word vectors. Such linear semantic composition is efficient, but does not always hold in human language. For example, the phrase “ivory tower”, which means “a state of privileged seclusion or separation from the facts and practicalities of the real world”, is not a linear combination of the individual meanings of “ivory” and “tower”. Instead, it carries a new meaning. We are therefore motivated to investigate a new language modeling paradigm to account for such intricate non-linear combination of word meanings.

Drawing inspiration from the recent findings in the emerging research area of quantum cognition, which suggest that human cognition [2–4, 14] especially language understanding [12, 13] exhibit certain non-classical phenomena (i.e. quantum-like phenomena), we propose a theoretical framework, named *Semantic Hilbert Space*, to formulate quantum-like phenomena in language understanding and to model different levels of semantic units in a unified space.

In Semantic Hilbert Space, we assume that words can be modeled as microscopic particles in superposition states, over the basic sememes (i.e. minimum semantic units in linguistics), while a combination of word meanings can be viewed as a mixed system of particles. The Semantic Hilbert Space represents different levels of semantic units, ranging from basic sememes, words and sentences, on a unified complex-valued vector space. In addition, we introduce a new semantic abstraction, named as *Semantic Measurements*, which are also embedded in the same vector space and trainable to extract high-level features from the mixed system.

As shown in Fig. 1, the Semantic Hilbert Space is built on the basis of quantum probability (QP), which is the probability theory for explaining the uncertainty of quantum superposition. As quantum superposition requires the use of the complex field, Semantic Hilbert Space has complex values and operators. In particular, the probability function is implemented by a (complex) density operator which is unique according to Gleason’s theorem. We argue that the complex-valued vector representation of words can inherently admit an mathematical formulation for complicated, non-linear combination of word meanings.

Semantic Hilbert Space adopts a complex-valued vector representation of unit length, where each component adopts an amplitude-phase form $z = re^{i\phi}$. We hereby hypothesize that the amplitude r and complex phase ϕ can be used to encode different levels of semantics such as lexical-level co-occurrence, hidden sentiment polarity or topic-level semantics. When word vectors are combined, even in a simple complex-valued addition form, the resulting expression will entail a non-linear composition of amplitudes and phases, thus indicating a complicated fusion of different levels of semantics. A more detailed explanation is given in Sec. 2. In this way, the complex-valued word embedding is fundamentally different from existing real-valued word embedding. A series of ablation tests indicate that the complex-valued word embedding can increase performance.

The Semantic Hilbert Space is an abstract representation of our approach to modeling language through QP. At the level of implementation, an efficient and effective computational framework is needed to cope with large text collections and user query streams. To do so, we propose an end-to-end neural network architecture, which provides means for training of the network components and modularity of components.

Each component corresponds to a physical meaning of quantum probability with well-defined mathematical constraints. Moreover, each component is easier to understand than the kernels in convolutional neural network and cells in recurrent neural networks. The network proposed in this paper is evaluated on six benchmarking datasets for text classification and achieves a steady increase over existing models. Moreover, it is shown that the proposed network is advantageous due to its high robustness and self-explanation capability.

2 SEMANTIC HILBERT SPACE

The mathematical foundation of Quantum Theory is established on a Hilbert Space over the complex field. In order to borrow the underlying mathematical formalism of quantum theory for language understanding, it is necessary to build such a Hilbert Space for language representation. In this study, we build a *Semantic Hilbert Space* $\mathcal{H}$ over the complex field. As is illustrated in Fig. 1, multiple levels of semantic units are modeled on this common Semantic Hilbert Space. In the rest of this section, the semantic units under modeling are introduced separately.

We follow the standard *Dirac Notation* for Quantum Theory. A unit vector $\vec{\mu}$ and its transpose $\vec{\mu}^T$ are denoted as a ket $|\mu\rangle$ and a bra $\langle\mu|$ respectively. The inner product and outer product of two unit vectors $\vec{u}$ and $\vec{v}$ are denoted as $\langle u|v\rangle$ and $|u\rangle\langle v|$ respectively.

2.1 Sememes

Sememes are the minimal non-separable semantic units of word meanings in language universals [20]. For example, the word "ironsmith" is composed of sememes "human", "occupation", "metal" and "industrial". We assume that the Semantic Hilbert Space $\mathcal{H}$ is spanned by a set of orthogonal basis $\{|e_j\rangle\}_{j=1}^n$ corresponding to a finite closed set of sememes $\{e_j\}_{j=1}^n$. In the quantum language, the set of sememes are modeled as *basis states*, which is the basis for representing any quantum state. In Fig. 1, the axes of the Semantic Hilbert Space correspond to the set of sememe states, and semantic units with larger granularity are represented on its basis.

2.2 Words

The meaning of a word is a combination of sememes. We adopt the concept of *superposition* to formulate this combination. Essentially, a word w is modeled as a quantum particle in *superposition state*, represented by a unit-length vector in the Semantic Hilbert Space $\mathcal{H}$, as can be seen in Fig. 1. It can be written as a linear combination of the basis states for sememes:

$$|w\rangle = \sum_{j=1}^{n} r_j e^{i\phi_j} |e_j\rangle \tag{1}$$

where the complex-valued weight $r_j e^{i\phi_j}$ denotes how much the meaning of word w is associated with the sememe e_j. Here $\{r_j\}_{j=1}^n$ 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. We could also transfer the complex number in a complex plane as $re^{i\phi} = r\cos\phi + ir\sin\phi$.

It is worth noting that the complex phases $\{\phi_j\}$ are crucial as they implicitly entail the *quantum interference* between words. Suppose two words w_1 and w_2 are of weights $r_j^{(1)} e^{i\phi_j^{(1)}}$ and $r_j^{(2)} e^{i\phi_j^{(2)}}$ for the sememe e_j. The two words in combination are therefore at the state e_j with a probability of

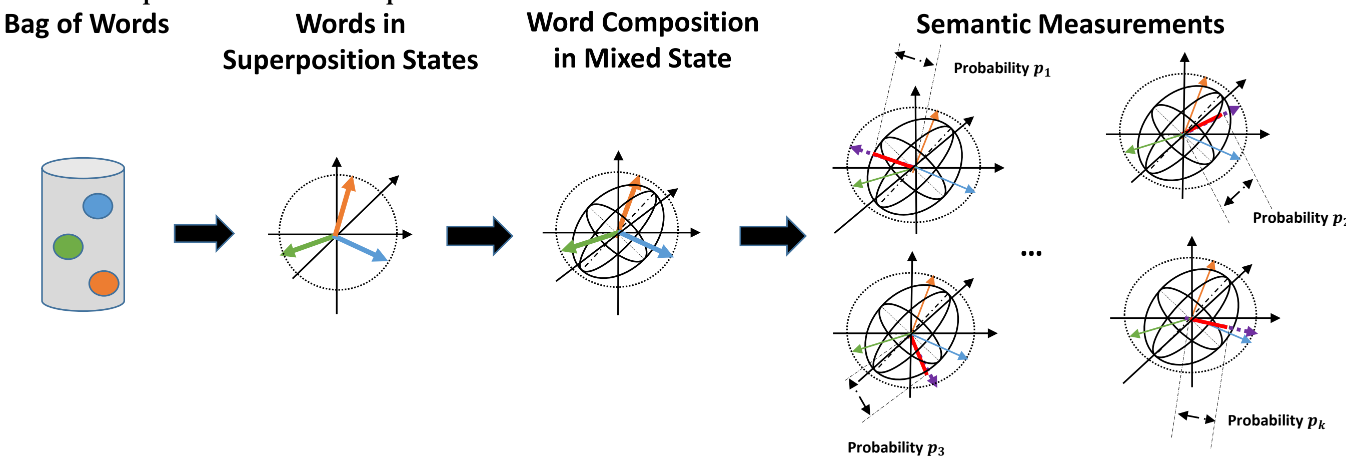
$$\begin{aligned} p &= |r_j^{(1)} e^{i\phi_j^{(1)}} + r_j^{(2)} e^{i\phi_j^{(2)}}|^2 \\ &= (r_j^{(1)})^2 + (r_j^{(2)})^2 + 2r_j^{(1)} r_j^{(2)} cos(\phi_j^{(1)} - \phi_j^{(2)}) \end{aligned} \tag{2}$$

Where the term $term2r_j^{(1)} r_j^{(2)} cos(\phi_j^{(1)} - \phi_j^{(2)}$ reflects the interference between the two words, where as the classical case corresponds to a particular case $\phi_j^{(1)} = \phi_j^{(2)} = 0$.

2.3 Semantic Compositions

As is illustrated in Fig. 1, we view a word composition (e.g. a sentence) as a bag of words [21], each of which is modeled as a particle in superposition state on the Semantic Hilbert Space $\mathcal{H}$. To obtain the semantic composition of words, we leverage the concept of *quantummixture* and formulate the

Figure 1: Illustration of Semantic Hilbert Space. The green, blue and orange colors correspond to three different words modeled as quantum particles. The black dotted circle represents the unit ball in the Semantic Hilbert Space. The ellipsoid in solid line refers to the quantum probability distribution defined by the density matrix of the word composition. The purple lines are semantic measurements. The intersections of the ellipsoids and semantic measurements are in thick red lines, the lengths of which correspond to measurement probabilities.



word composition as a mixed system composed of the word superposition states. The system is in a *mixed state* represented by a n-by-n density matrix ρ on $\mathcal{H}$, which is positive semi-definite with trace 1. It is computed as follows:

$$\rho = \sum_j p(j) \left| w_i \right\rangle \left\langle w_i \right|, \tag{3}$$

where $|w_i\rangle$ denotes the superposition state of the i^{th} word, is the classical probability of the state $|w_i\rangle$ with $\sum_j p(j) = 1$. It determines the contribution of the word w_i to the overall semantics.

The complex-valued density matrix ρ can be seen non-classical distribution of sememes in $\mathcal{H}$. Its diagonal elements are real and form a classical distribution of sememes, while its complex-valued off–diagonal entries encode the interplay between sememes, which in turn gives rise to the interference between words.

A density matrix assigns a probability value for any state on $\mathcal{H}$ such that the values for any set of orthogonal states sum up to 1 [19]. Hence it is be visualized as an ellipsoid in Fig. 1, the length of whose intersection with a unit vector denoting its quantum probability.

2.4 Semantic Measurements

As a non-classical probability distribution, a sentence density matrix carries rich and often redundant information. In order to extract the relevant information to a concrete task from the semantic composition, we build a set of measurements and compute the probability that the mixed system falls onto each of the measurements as a high-level abstraction of the semantic composition.

Suppose our proposed *semantic measurements* are associated with a set of measurement projectors $\{P_i\}_{i=1}^k$. According to the Born's rule [11], applying the measurement projector P_i onto the sentence density matrix ρ yields the following result:

$$p_i = tr(P_i \rho) \tag{4}$$

And the obtained probabilities $\{p_i\}_{i=1}^k$ are the high-level representation of ρ. Here, we only consider pure states as measurement states, i.e. $P_i = |v_i\rangle \langle v_i|$. Moreover, we ignore the constraints of the measurements states $\{|v_i\rangle\}_{i=1}^k$ (i.e. orthogonality or completeness) but keep them trainable, so that the most suitable measurements can be determined automatically by the data in a concrete task, such as classification or regression. In this way, the trainable semantic measurements can be understood as a similar approach to supervised dimensionality reduction [18], but in a quantum probability framework with complex values.

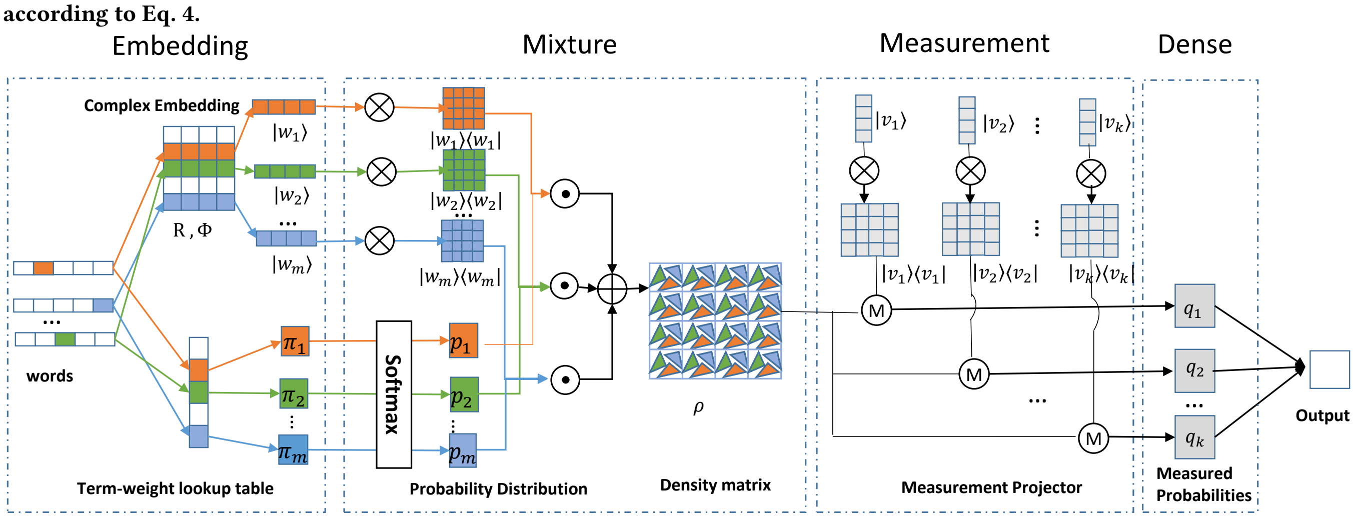
3 QUANTUM PROBABILITY DRIVEN NETWORK

In order to implement the proposed framework, we further propose an end-to-end neural network on its basis. Fig. 4 shows the architecture of the proposed Quantum Probability Driven Network (QPDN). In this section, we will introduce the architecture layer by layer.

3.1 Embedding Layer

The parameters of embedding Layer consist of $\{R, \Phi, \Pi\}$, denoting the amplitude embedding, phase embedding and

Figure 2: Architecture of Quantum probability-driven Neural Network. $\odot$ means that a matrix multiplies a number with each elements. $\bigoplus$ refers to a element-wise addition. $\bigotimes$ denotes a outer production to a vector, ⓜ means a measurement operation according to Eq. 4.



term-weight lookup table. Eq. 1 expresses a quantum representation as a unit-length, complex-valued vector representation for a word w, i.e. $|w\rangle = [r_1 e^{i\phi_1}, r_2 e^{i\phi_2} ... r_n e^{i\phi_n}]^T$. The term-weight lookup table is used to weight words for semantic combination, which will be described in the next subsection. During training, word embeddings need to be normalized to unit length after each batch. While it would be faster if we perform normalization after several batches [50].

3.2 Mixture Layer

A sentence is modeled as a density matrix, which is constructed in a bottom-up way in Sec. 2.3. Instead of using uniform weights in Eq. 3, word-sensitive weights are used for each words, which is commonly used in IR, e.g. inverse document frequency (IDF) as a word-dependent weight in TF-IDF scheme [42]. The new formula for the density matrix is given as follows:

$$\rho = \sum_{i}^{m} p(w_i) |w_i\rangle \langle w_i| \qquad (5)$$

In order to guarantee the unit trace length for density matrix, the word weights which are from the lookup table in a sentence are normalized to a probability value through a softmax operation: $p(w_i) = \frac{e^{\pi(w_i)}}{\sum_j^m e^{\pi(w_j)}}$. Compared to IDF weight, the normalized weight for a specific word in our approach is not static, but updated adaptively in training phase. Even in the inference/test phase, the real term weight i.e. $p(w_i)$ is also not static, but highly depends on the neighbor context words through nonlinear softmax function.

3.3 Measurement Layer

The measurement layer adopts a set of 1-order measurement projectors $\{|v_i\rangle \langle v_i|\}_{i=1}^k$, while $|v_i\rangle \langle v_i|$ is the outer product of its corresponding state in Semantic Hilbert Space $|v_i\rangle$. After each measurement, we can obtain a measured probability for each measurement state like $q_j = tr(\rho |v_j\rangle \langle v_j|)$. Finally, we can obtain a vector $\vec{q} = [q_1, q_2, ...q_k]$. Similarly to the word vectors which are also represented as unit states, the states $|v_i\rangle$ are also normalized after several batches.

3.4 Dense Layer

The vector $\vec{q}$ in measurement layer, which consists k positive scalar numbers, is used to infer the label for a given sentence. A dense layer with softmax activation is adopted after measurement layer to get a classification probability distribution, i.e. $\hat{\vec{y}} = \text{softmax}(\vec{q} \cdot W)$. The loss is designed as a cross-entropy loss between $\hat{\vec{y}}$ and the one-hot label $\vec{y}$.

4 EXPERIMENTS

4.1 Datasets

Our model is evaluated on 6 datasets for text classification: CR customer review [24], MPQA opinion polarity [48], SUBJ sentence subjectivity [37], MR movie review [37], SST binary sentiment classification [40], and TREC question classification [30]. The statistics of them are shown in Tab. 1.

4.2 Experimental Setup

4.2.1 Baselines. We compare the proposed QPDN with various models, including Uni-TFIDF, Word2vec, FastText [25] and Sent2Vec [36] as unsupervised representation learning baselines, CaptionRep [22] and DictRep [23] as supervised

Table 1: Dataset Statistics. (CV means 10-fold cross validation for testing performance.)

Dataset	train	test	vocab.	task	Classes
CR	4K	CV	6K	product reviews	2
MPQA	11k	CV	6K	opinion polarity	2
SUBJ	10k	CV	21k	subjectivity	2
MR	11.9k	CV	20k	movie reviews	2
SST	67k	2.2k	18k	movie reviews	2
TREC	5.4k	0.5k	10k	Question	6

Table 2: Experimental Results in percentage (%). The best performed value (except for CNN/LSTM) for each dataset is in bold. where † means a significant improvement over FasText.

Model	**CR**	**MPQA**	**MR**	**SST**	**SUBJ**	**TREC**
Uni-TFIDF	79.2	82.4	73.7	-	90.3	85.0
Word2vec	79.8	**88.3**	77.7	79.7	90.9	83.6
FastText [25]	78.9	87.4	76.5	78.8	91.6	81.8
Sent2Vec [36]	79.1	87.2	76.3	80.2	91.2	85.8
CaptionRep [22]	69.3	70.8	61.9	-	77.4	72.2
DictRep [23]	78.7	87.2	76.7	-	90.7	81.0
Ours: QPDN	**81.0**†	87.0	**80.1**†	**83.9**†	**92.7**†	**88.2**†
CNN [26]	81.5	89.4	81.1	88.1	93.6	92.4
BiLSTM [17]	81.3	88.7	77.5	80.7	89.6	85.2

representation learning baselines, as well as CNN [26] and BiLSTM [17] for advanced deep neural networks. We report the classification accuracy values of these models from the original papers.

4.2.2 Parameter Setting. In this paper, we use Glove word vectors [38] with 50,100,200 and 300 dimensions respectively. The amplitude embedding values are initialized by L2-norm, while the phases in complex-valued embedding are randomly initialized in $-\pi$ to π. We search for the best performance in a parameter pool, which contains a learning rate in {1e-3,1e-4,1e-5,1e-6}, an L2-regularization ratio in {1e-5,1e-6.1e-7,1e-8}, a batch size in {8,16,32,64,128}, and the number of measurements in {5,10,20,50,100,200}.

4.2.3 Parameter Scale. The main parameters in our model are amplitude embedding R and phase embedding Φ. Since both of them are $n \times |V|$ in shape, the number of parameters is roughly two times that of fastText [33]. For the other parameters, Π is $|V| \times 1$, $\{|v_i\rangle\}_{i=1}^{k}$ is $k \times 2n$, while W is $k \times |L|$ with L being the label set. Apart from word embeddings, the model is robust with limited scale at $k \times 2n + n \times |V| + k \times |L|$ for the number of parameters.

4.3 Results

The results in Tab. 1 demonstrate the effectiveness of our model, with improved classification accuracies over some strong baseline supervised and unsupervised representation models on most of the datasets except MPQA. In comparison with more advanced models including BiLSTM and CNN, our model generally performs better than BiLSTM with increased accuracy values on the multi-class classification dataset (TREC) and three binary text classification datasets (MR, SST & SUBJ). However, it under-performs CNN on all 6 datasets with a difference of over 2% on 3 of them (MPQA, SST & TREC), probably because that it uses fewer parameters and simpler structures. We argue that QPDN achieves a good balance between effectiveness and efficiency, due to the fact that it outperforms BiLSTM.

Table 3: Physical meanings and constraints

Components	DNN	QPDN
Sememe	-	basis vector / **basis state** $\{w \mid w \in C^n, \|\|w\|\|_2 = 1, \}$ complete &orthogonal
Word	real vector $(-\infty, \infty)$	unit complex vector / **superposition state** $\{w \mid w \in C^n, \|\|w\|\|_2 = 1\}$
Low-level representation	real vector $(-\infty, \infty)$	density matrix / **mixed system** $\{\rho \mid \rho = \rho^*, tr(\rho) = 1\}$
Abstraction	CNN/RNN $(-\infty, \infty)$	unit complex vector / **measurement** $\{w \mid w \in C^n, \|\|w\|\|_2 = 1\}$
High-level representation	real vector $(-\infty, \infty)$	probabilities/ **measured probability** (0, 1)

5 DISCUSSIONS

This section discusses on the power of self-explanation and conducts an ablation test to examine the usefulness of important components of the network, especially the complex-valued word embedding. Additionally , the intuitive meaning of the discriminative semantic directions is shown in Sec. 5.3.

5.1 Self-explanation Components

As is shown in Tab. 3, all components in our model have a clear physical meaning corresponding to quantum probability, where classical Deep Neural Network (DNN) can not well explain the role each component plays in the network. Essentially, we construct a bottom-up framework to represent each level of semantic units on a uniform Semantic Hilbert Space, from the minimum semantic unit, i.e. sememe, to the sentence representation. The framework is operationalized through superposition, mixture and semantic measurements. On the one hand, the explanation is reflected by well-designed constraints for all the components. On the other hand, some intuitive explanation can be performed on the crucial components of the network i.e, measurements, as shown in Sec. 5.3.

Table 4: Ablation Test

Setting	SST	Δ
FastText [25]	0.7880	-0.0511
FastText [25] with double-dimension real word vectors	0.7883	-0.0508
fixed amplitude part but trainable phase part	0.8199	-0.0192
replace trainable weights with fixed mean weights	0.8303	-0.0088
replace trainable weights with fixed IDF weights	0.8259	-0.0132
non-trainable projectors with fixed orthogonal ones	0.8171	-0.0220
replace projectors with dense layer	0.8221	-0.0170
QPDN	**0.8391**	-

Table 5: The learned measurement for dataset MR. They are selected according to nearest words for a measurement vector in Semantic Hibert Space

Measurement	Selected neighborhood words
1	change, months, upscale, recently, aftermath
2	compelled, promised, conspire, convince, trusting
3	goo, vez, errol, esperanza, ana
4	ice, heal, blessedly, sustains, make
5	continue, warned, preposterousness, adding, falseness

5.2 Ablation Test

An ablation test is conducted to examine how each component influences the final performance of QPDN. In particular, a double-length real word embedding network is implemented to examine the use of complex-valued word embedding, while mean weights and IDF weights are used as alternative word weighting strategies to check the necessity of introducing trainable weights. A set of non-trainable orthogonal projectors and a dense layer on top of the sentence density matrix are implemented to analyze the effect of trainable semantic measurements.

We use 100-dimensional real-valued word vectors and 50-dimensional complex-valued vectors for the models in the ablation test. All models under ablation are comparable in terms of time cost. Tab. 4 shows that each component plays an important role in the QPDN model. In particular, replacing complex embedding with double-dimension real word embedding leads to a 5% drop in performance, which indicates that the complex-valued word embedding is not merely doubling the number of parameters.

5.3 Discriminative Semantic Directions

In order to better understand the well-trained measurement projectors, we obtained the top 10 nearest words in complex-valued vector for each trained measurement state (like $|v_i\rangle$), using KD tree [9]. Due to limited space, we take 5 measurements from the trained model for the MR dataset, and select words from the top 10 nearest words to each measurement. As can be seen in Tab. 5, the first measurement is roughly about changes over time, the second concerning being motivated or forced to do something. While the third measurement groups uncommon non-English words together. The last two measurement also group words sharing similar meanings. It is therefore interesting to see that relevant words can somehow be grouped together into certain topics during the training process, which may be discriminative for the given task.

6 RELATED WORKS

6.1 Word embedding based Sentence modelling

Neural language model is firstly proposed to use neural network for language modelling, and it has a side effect for word embedding, and the it is further investigated by [8] and further developed by [15, 34], tends to be more and more dominant. Word2vec [33] makes the training of word embedding in more fast way e.g. CBow and Skip-gram, with removing non-linear layers and other tricks e.g. hierarchical softmax and negative sampling. The another popular word embedding named Glove [38] make advantage of global matrix factorization and count-based methods. However, we argue that it is hard to represent the complex meaning of a word with only a real-valued vector (i.e. a point in a low-dimension space), especially in the cases of polysemy or ambiguity. Gaussian Embedding [46] is a alternative density-based approach to represent a word as a Guassian distribution. This inspires us to exploit new word representation, namely complex-valued embedding for language modelling.

As the mainstream approach, word embedding mainly models a word as a fix-length vector to represent its semantics, and combines word embeddings to obtain sentence representation. It can be achieved through unsupervised [22, 27, 35], supervised [5, 6] approaches or by deep deep neural networks [26, 43]. Essentially, these works leverage lexical-level co-occurrences or sequential orders as the basis of word semantic features, and apply highly linear operations to word features in order to obtain a sentence representation. Even though they have achieved superior performances in many NLP tasks, they are insufficient to explicitly model the more complicated non-linear semantic combination of words as described above. Meanwhile, most of current models are hard to be understood by end users, from both intuitive and theoretical point of view.

6.2 Quantum Theory in Information Retrieval

Ever since the pioneering work of Van Rijsbergen [45], Quantum Theory has been applied in Information Retrieval, yielding a series of novel retrieval approaches, such as quantum probability ranking principle [39, 53] and quantum language model [29, 41, 49]. However, these research are limited to ad-hoc retrieval scenarios and fail to address more complicated tasks in information access and retrieval (IAR), which requires fine-grained modeling of textual data. In such scenarios as text classification, question answering, reading comprehension and multi-turn dialogue systems, embedding-based neural networks are the state of art. The first quantum-inspired model specifically aiming at these tasks is a Neural Network based Quantum-like Language Model (NNQLM) [51]. Nonetheless, it is still a traditional neural network with real-valued density matrices and CNN, and therefore insufficient to capture the essential properties of quantum mechanics, with only real-valued represenations. To the best of our knowledge, our study is the first to propose a unified quantum probability framework for fundamental text understanding tasks, with complex-valued representations.

6.3 Quantum-like Phenomena for Language

Quantum-like phenomena in human cognition of information[2–4, 13, 14], especially in language understanding, has recently been investigated [14]. For example, Bruza et al. [13] discussed word association/entanglement and pointed out that a context might affect word association via the standard interpretation of quantum measurement. Wang et al. [47] conducted a series of user studies to demonstrate that the document relevance judgment could be largely affected by the previous documents in a quantum-like interference manner. Blacoe et al. [10] proposed a bottom-up quantum approach to model distributed semantics with dependency graphs, achieving a competitive performance in word similarity and association tasks. T Basile and Tamburini [7] build a new language model based on quantum notation with state evolution. These works provide us with inspirations and insights to investigate in a quantum probabilistic framework to enhance language understanding.

6.4 Other related Approaches

The discussions on the use of complex numbers have started as early as the advent of quantum-inspired information retrieval (QIR) models. Van Rijsbergen, in his pioneering book of QIR [45], suggested to use complex numbers as a mechanism to store information, and proposed to assign the amplitude and complex phase to represent TF (Term Frequency) and IDF (Inverse Document Frequency) of a term respectively. This idea was further discussed in Zuccon et al. [54] and implemented on real ad-hoc retrieval datasets. However, it is still a simple and empirical attempt. Recently, a deep complex network [44] was proposed, where all real layers were replaced with their complex-valued counterparts. Despite the complex-valued structure of the network, the inputs and outputs are real-valued vectors like in any general neural networks, and the model is not related to quantum probability. Aerts et al. [1] elaborated a quantum model for the concepts in a collection of web documents, where a concept admits a complex-valued representation containing an amplitude and a complex phase, but with only simple illustrative examples. Lin et al. [31] used amplitude and phase as the parameters to implement the inference process with diffraction principle, lead to a light-speed prediction for image classification. This also provides a very huge potentials for our QPDN in the language with light-speed inference.

Levine et al. [28] attempted to draw a connection between quantum entanglement and neural networks, and illustrated that neural networks and quantum mechanics can be more closely connected and integrated. It also motivates us to carry out deeper investigations on this topic in the current and future studies. Zhang et al. [52] proposes a CNN implement for the decomposition of quantum many body function, which reveals the potential of quantum insights from empirical point of view. Liu et al. [32] proposed a novel text classifier with training the Hamiltonian matrix and the unitary operator for better fitting the dataset. This model preliminarily reveals the potential of the quantum classifier in supervised learning tasks.

7 CONCLUSIONS

In order to better model the non-linearity of word semantic composition, we have developed a quantum-inspired framework that models different granularities of semantic units on the same Semantic Hilbert Space, and implement this framework into an end-to-end text classification network. The network shows a promising performance on 6 benchmarking text datasets, in effectiveness, robustness and self-explanation ability. Moreover, the complex-valued word embedding approach, which inherently achieves non-linear combination of word meanings, does bring benefits to the classification accuracy in a comprehensive ablation study.

This work is among the first steps to apply the quantum probabilistic framework to text modeling. We believe it is a promising direction. On one hand, we would like to further extend this work by considering deeper and more complicated structures such as attention or memory mechanism in language, in order to investigate related quantum-like phenomena on textual data to provide more intuitive insights.

ACKNOWLEDGMENTS

This work is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 721321.

REFERENCES

[1] Diederik Aerts, Jonito Aerts Arguelles, Lester Beltran, Lyneth Beltran, Isaac Distrito, Massimiliano Sassoli de Bianchi, Sandro Sozzo, and Tomas Veloz. 2017. Towards a Quantum World Wide Web. *arXiv:1703.06642 [quant-ph]* (March 2017). http://arxiv.org/abs/1703.06642 arXiv: 1703.06642.

[2] Diederik Aerts and Marek Czachor. 2004. Quantum Aspects of Semantic Analysis and Symbolic Artificial Intelligence. *Journal of Physics A: Mathematical and General* 37, 12 (March 2004), L123–L132. https://doi.org/10.1088/0305-4470/37/12/L01 arXiv: quant-ph/0309022.

[3] Diederik Aerts, Liane Gabora, and Sandro Sozzo. 2013. Concepts and Their Dynamics: A Quantum-Theoretic Modeling of Human Thought. *Topics in Cognitive Science* (Sept. 2013). https://doi.org/10.1111/tops.12042 arXiv: 1206.1069.

[4] Diederik Aerts and Sandro Sozzo. 2014. Quantum Entanglement in Concept Combinations. *International Journal of Theoretical Physics* 53, 10 (Oct. 2014), 3587–3603. https://doi.org/10.1007/s10773-013-1946-z

[5] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. (Nov. 2016). https://openreview.net/forum?id=SyK00v5xx

[6] John Wieting Mohit Bansal and Kevin Gimpel Karen Livescu. 2016. TOWARDS UNIVERSAL PARAPHRASTIC SENTENCE EMBEDDINGS. (2016), 19.

[7] Ivano Basile and Fabio Tamburini. 2017. Towards quantum language models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1840–1849.

[8] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *JMLR* 3, Feb (2003), 1137–1155.

[9] Jon Louis Bentley. 1975. Multidimensional binary search trees used for associative searching. *Commun. ACM* 18, 9 (1975), 509–517.

[10] William Blacoe, Elham Kashefi, and Mirella Lapata. 2013. A quantum-theoretic approach to distributional semantics. In *NAACL*. 847–857.

[11] Max Born. 1926. Zur Quantenmechanik der Sto\s svorgÃďnge. *Zeitschrift fÃijr Physik* 37, 12 (Dec. 1926), 863–867. https://doi.org/10.1007/BF01397477

[12] Peter Bruza, Kirsty Kitto, Douglas Nelson, and Cathy McEvoy. 2009. Is there something quantum-like about the human mental lexicon? *Journal of Mathematical Psychology* 53, 5 (2009), 362–377.

[13] Peter D Bruza, Kirsty Kitto, Douglas McEvoy, and Cathy McEvoy. 2008. Entangling words and meaning. (2008).

[14] Jerome R Busemeyer and Peter D Bruza. 2012. *Quantum models of cognition and decision*. Cambridge University Press.

[15] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*. ACM, 160–167.

[16] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *EMNLP*. Association for Computational Linguistics, 670–680. http://aclweb.org/anthology/D17-1070

[17] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *arXiv:1705.02364 [cs]* (May 2017). arXiv: 1705.02364.

[18] Ronald A Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics* 7, 2 (1936), 179–188.

[19] Andrew M Gleason. 1957. Measures on the closed subspaces of a Hilbert space. *Journal of mathematics and mechanics* (1957), 885–893.

[20] Cliff Goddard and Anna Wierzbicka. 1994. *Semantic and lexical universals: Theory and empirical findings*. Vol. 25. John Benjamins Publishing.

[21] Zellig S Harris. 1954. Distributional structure. *Word* 10, 2-3 (1954), 146–162.

[22] Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning Distributed Representations of Sentences from Unlabelled Data. In *NAACL*. Association for Computational Linguistics, San Diego, California, 1367–1377. http://www.aclweb.org/anthology/N16-1162

[23] Felix Hill, KyungHyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to Understand Phrases by Embedding the Dictionary. *TACL* 4 (2016), 17–30. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/\article/view/711

[24] Minqing Hu and Bing Liu. 2014. Mining and Summarizing Customer Reviews. (2014), 10.

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

[26] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. *EMNLP* (2014), 1746–1751. https://doi.org/10.3115/v1/D14-1181

[27] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought Vectors. In *NIPS (NIPS'15)*. MIT Press, Cambridge, MA, USA, 3294–3302.

[28] Yoav Levine, David Yakira, Nadav Cohen, and Amnon Shashua. 2017. Deep learning and quantum entanglement: Fundamental connections with implications to network design. *arXiv preprint arXiv:1704.01552* (2017).

[29] Qiuchi Li, Jingfei Li, Peng Zhang, and Dawei Song. 2015. Modeling multi-query retrieval tasks using density matrix transformation. In *SIGIR*. ACM, 871–874.

[30] Xin Li and Dan Roth. 2002. Learning Question Classifiers. In *COLING*. Association for Computational Linguistics.

[31] Xing Lin, Yair Rivenson, Nezih T Yardimci, Muhammed Veli, Yi Luo, Mona Jarrahi, and Aydogan Ozcan. 2018. All-optical machine learning using diffractive deep neural networks. *Science* 361, 6406 (2018), 1004–1008.

[32] Ding Liu, Xiaofang Yang, and Minghu Jiang. 2013. A Novel Classifier Based on Quantum Computation. In *ACL*, Vol. 2. 484–488.

[33] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[34] Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on Machine learning*. ACM, 641–648.

[35] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2017. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. *arXiv:1703.02507 [cs]* (March 2017). arXiv: 1703.02507.

[36] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. *NAACL* 1 (2018), 528–540. https://doi.org/10.18653/v1/N18-1049

[37] Bo Pang and Lillian Lee. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. Association for Computational Linguistics, 115–124. https://doi.org/10.3115/1219840.1219855

[38] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation.. In *EMNLP*, Vol. 14. 1532–1543.

[39] Emmanuel M Pothos and Jerome R Busemeyer. 2009. A quantum probability explanation for violations of 'rational' decision theory. *Proceedings of the Royal Society of London B: Biological Sciences* (2009), rspb–2009.

[40] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *EMNLP* (2013), 1631–1642.

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

[42] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* 28, 1 (1972), 11–21.

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

[44] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, João Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal. 2017. Deep complex networks. *arXiv preprint arXiv:1705.09792* (2017).

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

[46] Luke Vilnis and Andrew McCallum. 2014. Word representations via gaussian embedding. *arXiv preprint arXiv:1412.6623* (2014).

[47] Benyou Wang, Peng Zhang, Jingfei Li, Dawei Song, Yuexian Hou, and Zhenguo Shang. 2016. Exploration of quantum interference in document relevance judgement discrepancy. *Entropy* 18, 4 (2016), 144.

[48] Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation* 39, 2-3 (May 2005), 165–210. https://doi.org/10.1007/s10579-005-7880-9

[49] Mengjiao Xie, Yuexian Hou, Peng Zhang, Jingfei Li, Wenjie Li, and Dawei Song. 2015. Modeling quantum entanglements in quantum language models. (2015).

[50] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *NAACL*. 1006–1011.

[51] Peng Zhang, Jiabin Niu, Zhan Su, Benyou Wang, Liqun Ma, and Dawei Song. 2018. End-to-End Quantum-like Language Models with Application to Question Answering. (2018).

[52] Peng Zhang, Zhan Su, Lipeng Zhang, Benyou Wang, and Dawei Song. 2018. A Quantum Many-body Wave Function Inspired Language Modeling Approach. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 1303–1312.

[53] Guido Zuccon and Leif Azzopardi. 2010. Using the quantum probability ranking principle to rank interdependent documents. In *ECIR*. Springer, 357–369.

[54] Guido Zuccon, Benjamin Piwowarski, and Leif Azzopardi. 2011. On the use of complex numbers in quantum models for information retrieval. In *ICTIR*. Springer, 346–350.