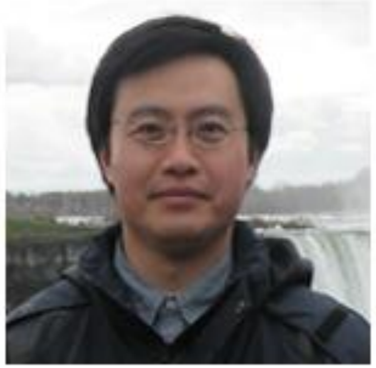
Exploring **Interpretable** Quantum Representation for language understanding

Benyou Wang, Qiuchi Li, Prayag Tiwari, Massimo Melucci

University of Padova

18/sep/2018

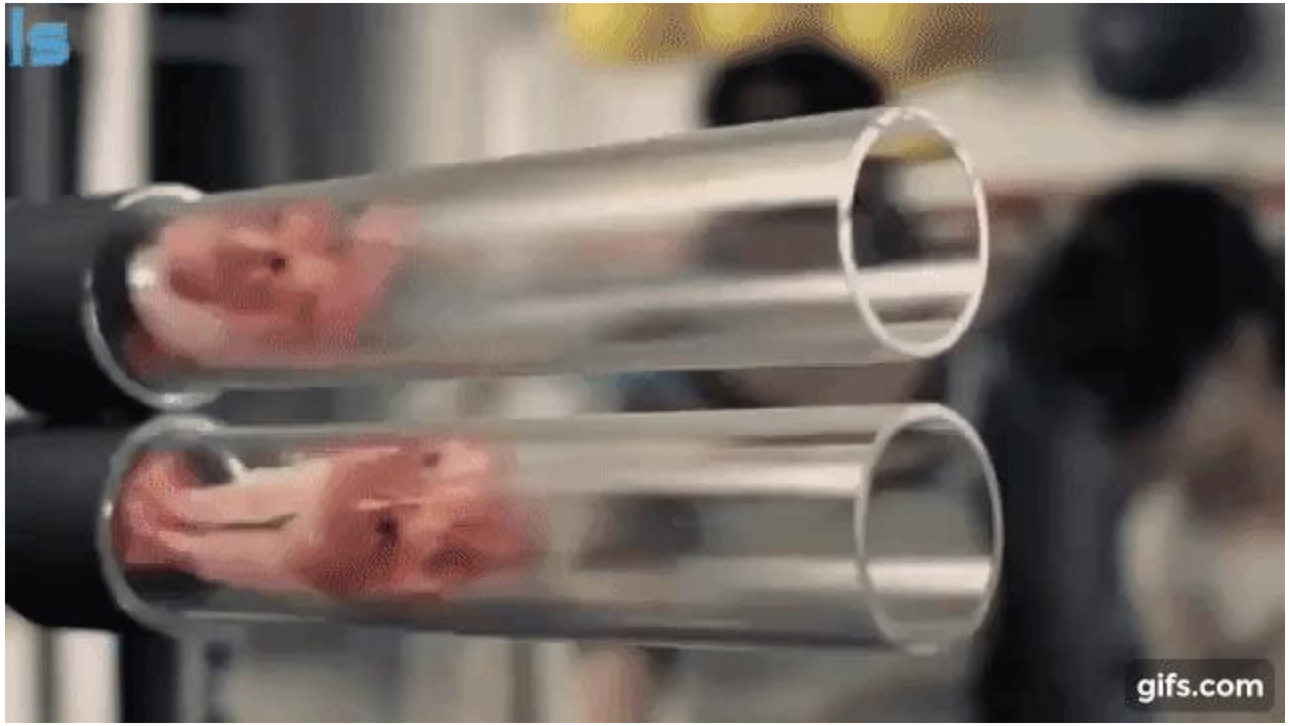
Done with the collaboration



Contents

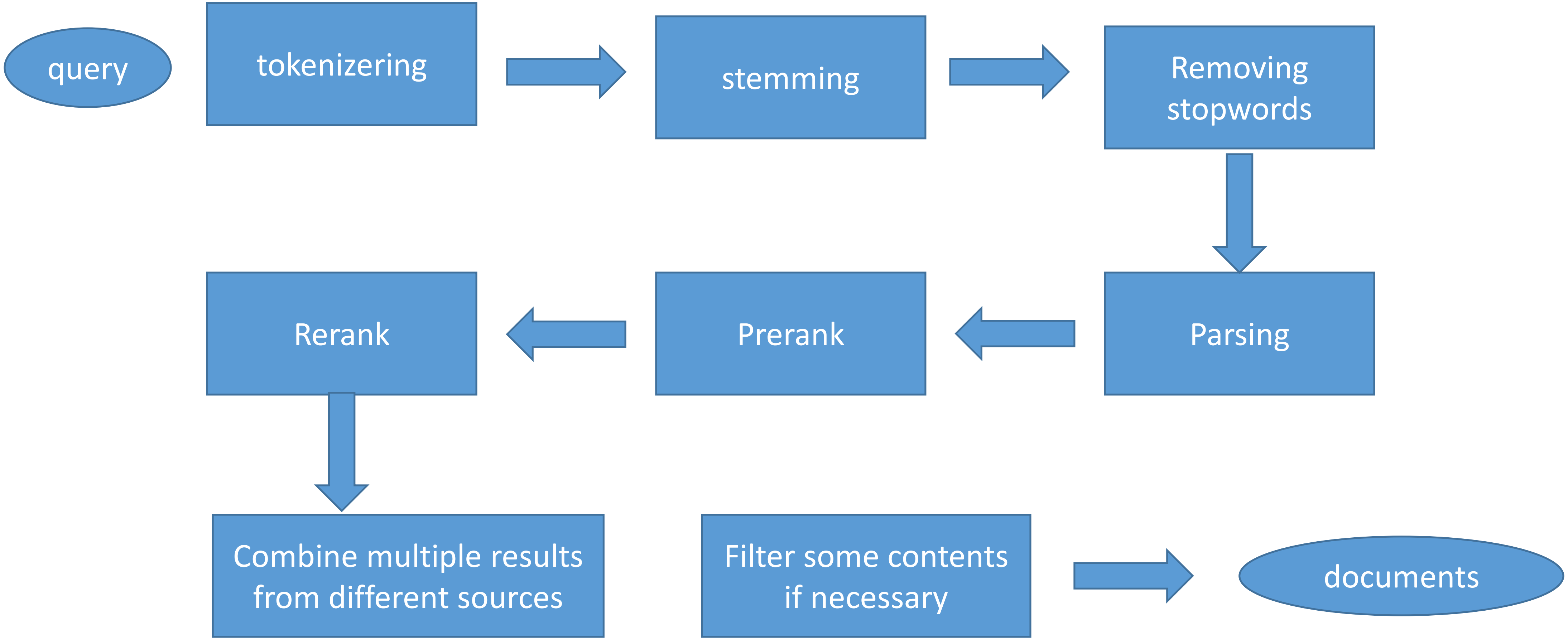
- Motivation: Interpretability in end2end network
- Method: Hilbert Semantic Space
- Applications: language representation and matching
  - Text classification
  - Matching with question answering

Transparency in end-to-end Paradigm



https://www.youtube.com/watch?v=TYpBJ71VW9g

An **Pipeline** example for text processing



End to end mechanism

- ✓ Less accumulating error
- ✓ Less involvement with Human beings
- ✓ Improve performance with shared features of the downstream tasks and upstream tasks

- ❖ Hard to adjust
- ❖ Hard to transfer
- ❖ Hard to understand

We need End to End mechanism, but in a fine-grained way

What is **Interpretability**

- Post-hoc explanations
  - Take a **learned model** and draw some kind of useful insights
  - E.g. Visualization in machine translation [Liu Yang &Maosong Sun ACL 2017]

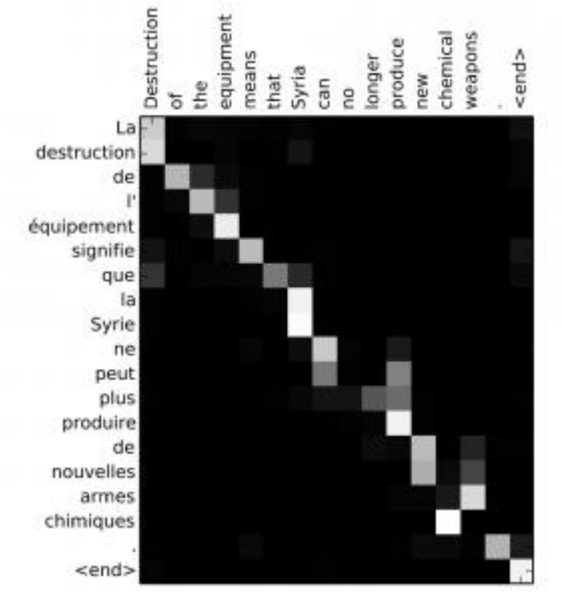
- Transparency
  - Targeting ``how does the model work?'' and seeks to provide some way to understand the core mechanisms
  - E.g. Capsule Network [Hinton NIPS 2017]

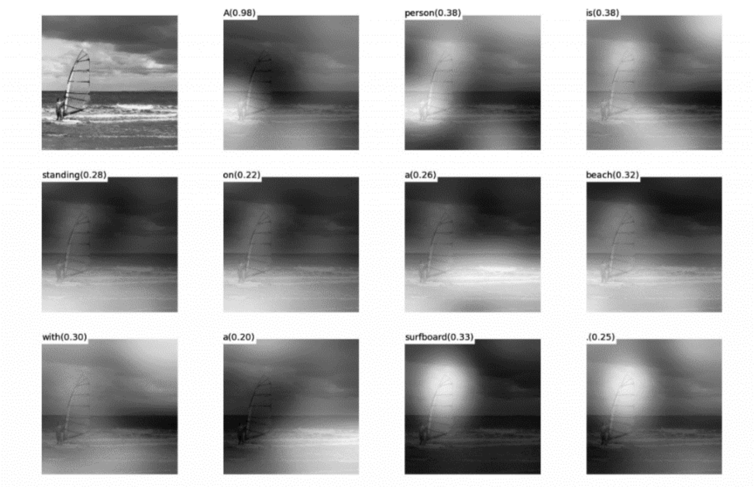
Zachary C Lipton. The mythos of model interpretability. arXiv preprint arXiv:1606.03490, 2016， **ICML** Workshop on Human Interpretability in Machine Learning
Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. Visualizing and understanding neural machine translation. **ACL**, volume 1, pages 1150–1159, 2017.
Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules[C]//**NIPS . 2017**: 3856-3866.

Interpretability: **Attention**

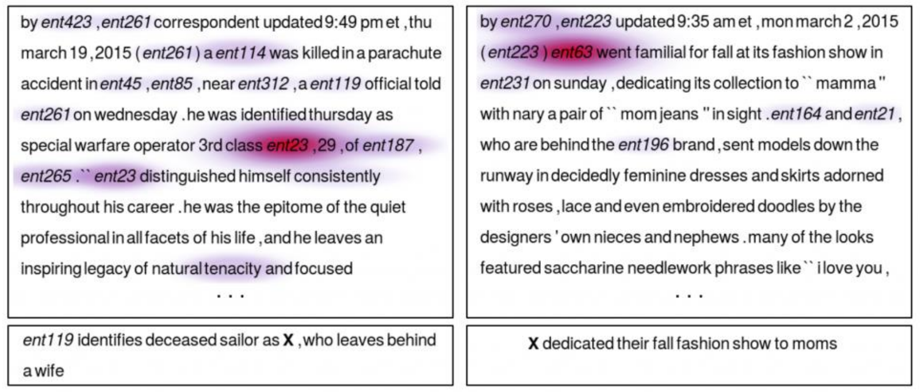
For a given vector $\vec{w}$, we normalize it with **softmax** thus guarantee their sum equals to 0

$$\vec{w}' = softmax(\vec{w}), \;\; w_i = \frac{e^{w_i}}{\sum e^{w_i}}$$





(b) A person is standing on a beach with a surfboard.



Design each subcomponents in the End-2-end architecture with a good background of the task

- *Both language understanding and artificial intelligence require being able to understand bigger things from knowing about smaller parts*

Christopher Manning 2017

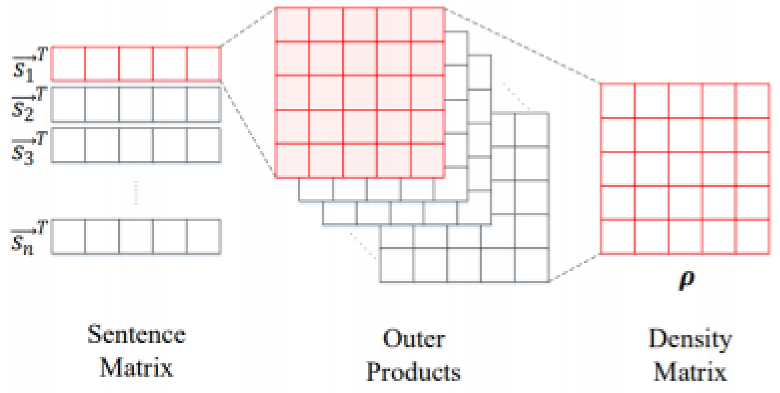
Motivations

- Design self-***explainable*** subcomponents in end2end network
- Provides more **transparency** as well as **Post-hoc explanations**
- **Theoretically-sound** network

Related works

- End to End language model for QA [AAAI 2018]
- Quantum Many body function for language model in QA **[CIKM 2018]**
- Quantum-inspired word Embedding [ACL REP4NLP 2018]
- **Hibert Semantic Space [In process]**

End-2-end Language model for QA



Matching with two matrices

- $tr(\rho_1 \rho_2)$
- CNN over $\rho_1 \rho_2$

Zhang Peng, Niu Jiabing, Su Zhan, Wang Benyou et al. End-to-End Quantum-like Language Models with Application to Question Answering **AAAI 2018**

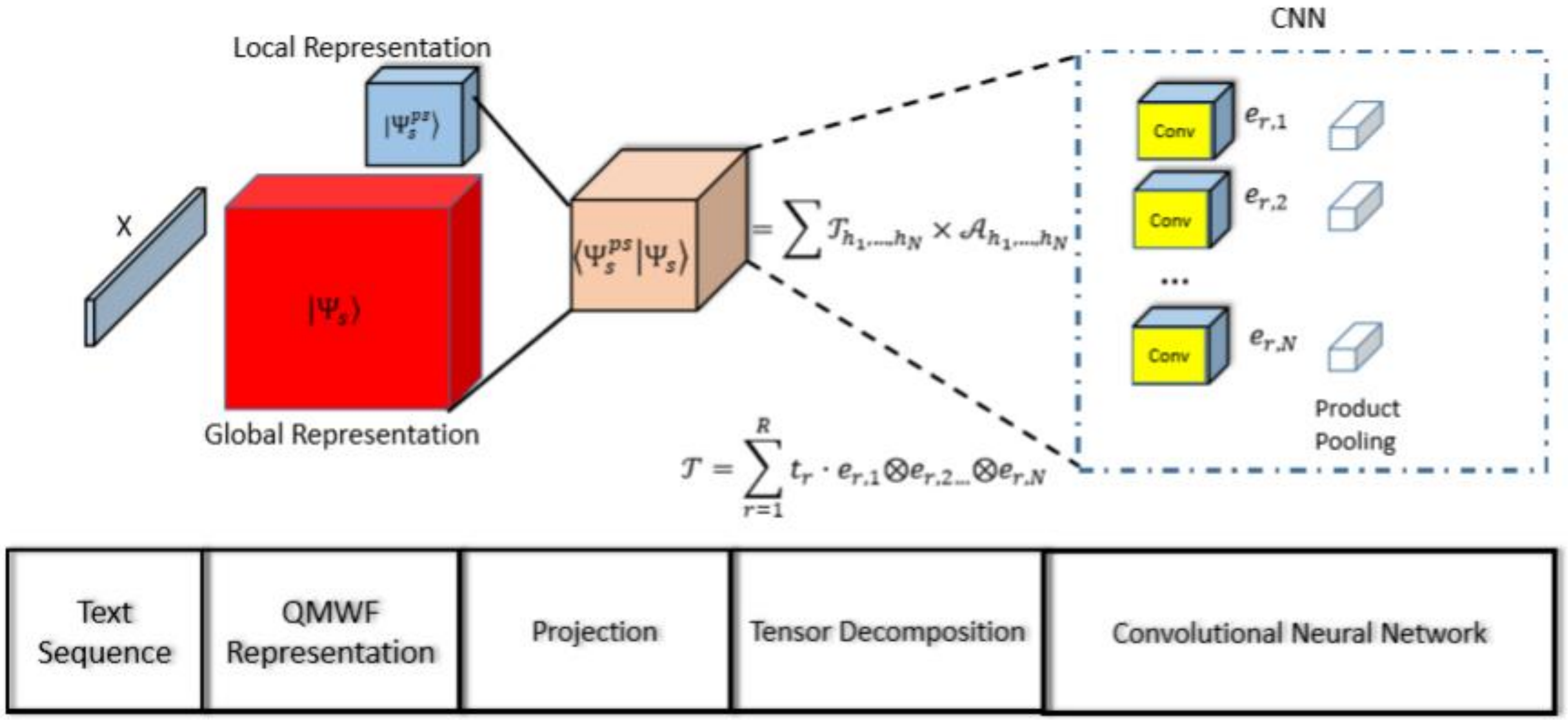
Metric/similarity for $\rho_q \rho_a$ [e.g. tr($\boldsymbol{\rho_q \rho_a}$) or $\boldsymbol{f}_{\mathbf{cnn}}(\boldsymbol{\rho_q \rho_a})$]

- Not theoretically-sound
  - $tr(\rho_q \rho_a)$ can not obtain the maximum value if $\rho_q \neq \rho_a$
  - Can not guarantee $tr(\rho_q \rho_x) + tr(\rho_x \rho_a) > tr(\rho_q \rho_a)$

- Ignoring the mathematical property of density matrix (probability distribution)

- Others
  - Real-valued based instead of complex-valued
  - Can not guarantee the unity length of density matrix.

Quantum many-body function for LM



Use CNN to ***approximate*** Tensor Decomposition in the projection of Quantum Many-Body Language Function

Peng Zhang, Zhan Su, Lipeng Zhang, Benyou Wang , Dawei Song. 2018. A Quantum Many-body Wave Function Inspired Language Modeling Approach, **CIKM 2018**

Complex word-embedding

- Super-linearity superposition with phase

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

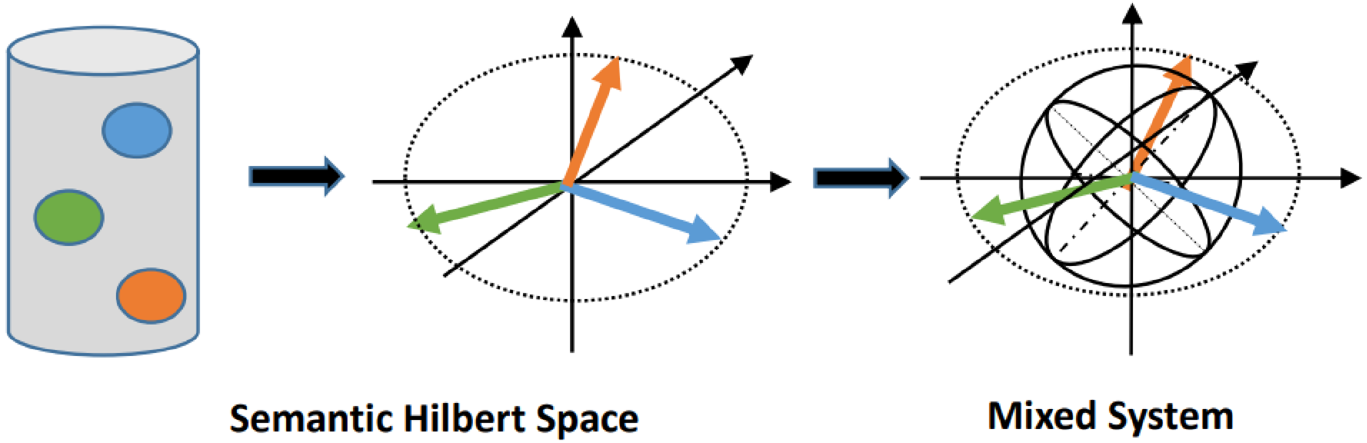
Li Qiuchi, Uprety Sagar, Wang Benyou , Song Dawei Quantum-inspired Complex Word Embedding, ACL 2018 3rd Workshop on Representation Learning for NLP , **ACL 2018 RepL4NLP**

Hilbert Semantic Space

- **Unify** these four things in a complex-valued space
  - Sememes
  - Word
  - Phrase/Sentence/Documents
  - Topic as measurements

Definition

- Sememes as basic state
- Word as superstition state
- Sentence as mixed system

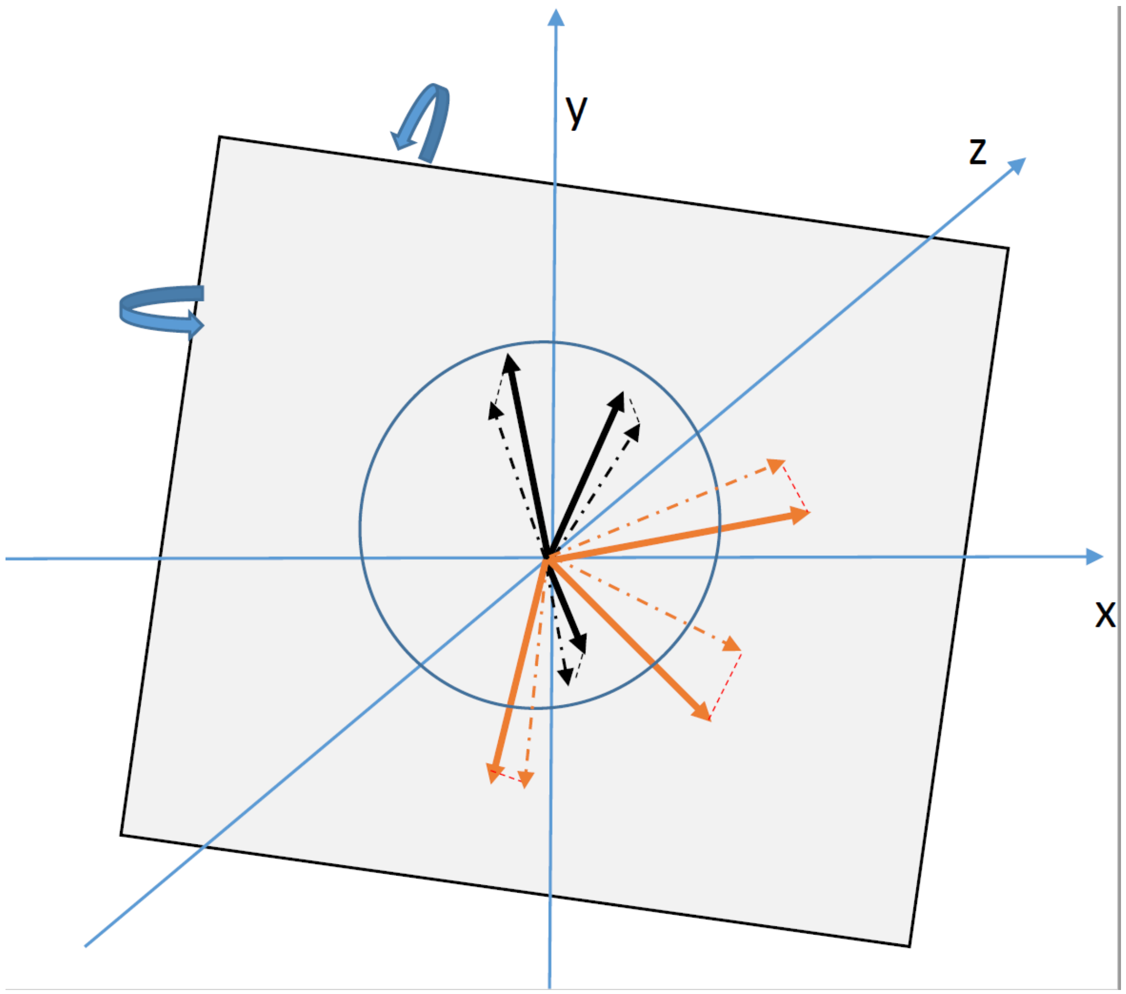


Complex word embedding

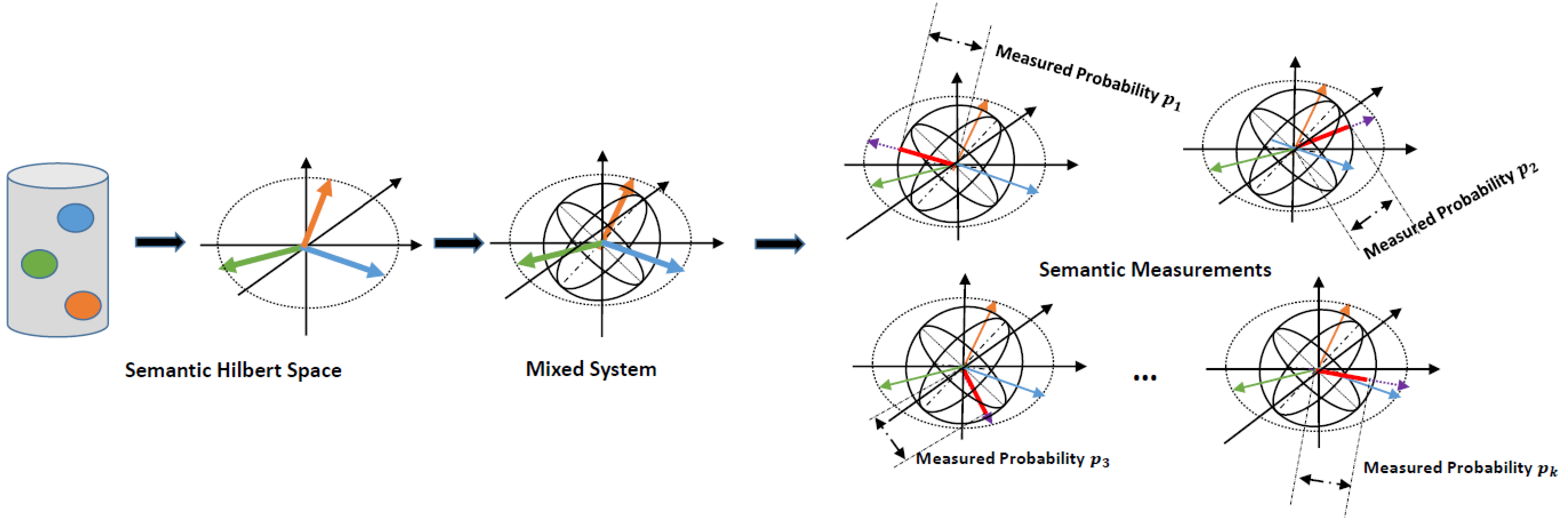
- Dimension: the number of
- Length : weight
- Amplitude part: meaning
- Phase part: polarity ?

- How to infer the overall polarity from the polarity of each words?
  - Is there any quantum phenomena here ?

Trainable Measurements for sentence classification

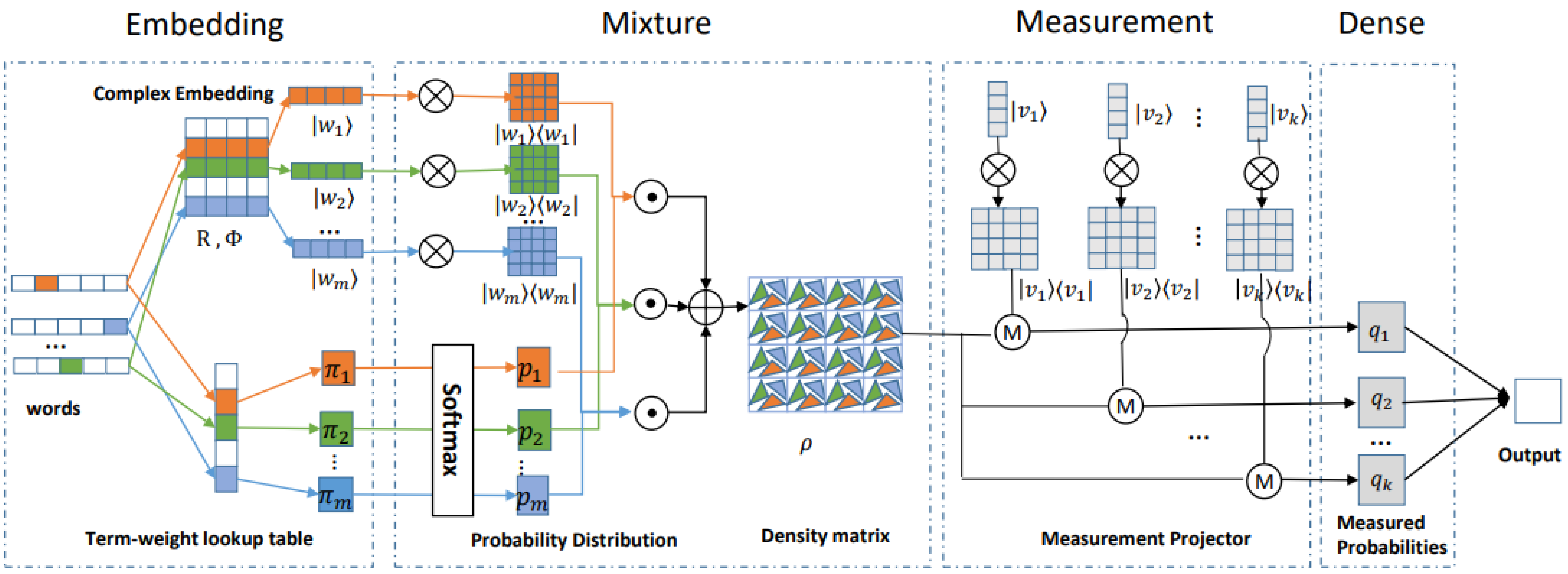


Framework



Semantic Hilbert Space
Mixed System
Measured Probability p_1
Measured Probability p_2
Semantic Measurements
Measured Probability p_3
...
Measured Probability p_k

Implements



Embedding
Mixture
Measurement
Dense
Complex Embedding
$|w_1\rangle$
$|w_2\rangle$
$|w_m\rangle$
R , Φ
words
Term-weight lookup table
π_1
π_2
π_m
Softmax
p_1
p_2
p_m
$|w_1\rangle\langle w_1|$
$|w_2\rangle\langle w_2|$
$|w_m\rangle\langle w_m|$
Probability Distribution
ρ
Density matrix
$|v_1\rangle$
$|v_2\rangle$
$|v_k\rangle$
$|v_1\rangle\langle v_1|$
$|v_2\rangle\langle v_2|$
$|v_k\rangle\langle v_k|$
M
Measurement Projector
q_1
q_2
q_k
Measured Probabilities
Output

Physical meaning for our models

Table 3: Physical meaning and constraint for each component

Components	Traditional DNN	NNQLM [56]	QPDN
Input embedding	arbitrary real vector $(-\infty, \infty)$	arbitrary real vector $(-\infty, \infty)$	unit complex vector, corresponding to superposition state $\{w \mid w \in C^n, \|\|w\|\|_2 = 1\}$
Low-level representation	arbitrary real vector $(-\infty, \infty)$	fake, real-valued density matrix $\{\rho \mid \rho \in \mathcal{R}^{n*n}\}$,	density matrix, corresponding to mixed state $\{\rho \mid \rho = \rho^*, tr(\rho) = 1, \mu\rho\mu^T > 0 \ \forall \mu \neq \overrightarrow{0}, \rho \in C^{n*n}\}$,
Abstraction	CNN/RNN/Attention $(-\infty, \infty)$	CNN $(-\infty, \infty)$	measurement vector, corresponding to measurement $\{w \mid w \in C^n, \|\|w\|\|_2 = 1\}$
High-level representation	arbitrary real vector $(-\infty, \infty)$	arbitrary real vector $(-\infty, \infty)$	real-valued probability, corresponding to measurement result $(0, 1)$

Experiments

Table 2: Experiment Results in percentage(%). The best performed value (except for CNN/LSTM) for each dataset is in bold.

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 [28]	78.9	87.4	76.5	78.8	91.6	81.8
Sent2Vec [42]	79.1	87.2	76.3	80.2	91.2	85.8
CaptionRep [21]	69.3	70.8	61.9	-	77.4	72.2
DictRep [22]	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 [29]	81.5	89.4	81.1	88.1	93.6	92.4
BiLSTM [16]	81.3	88.7	77.5	80.7	89.6	85.2

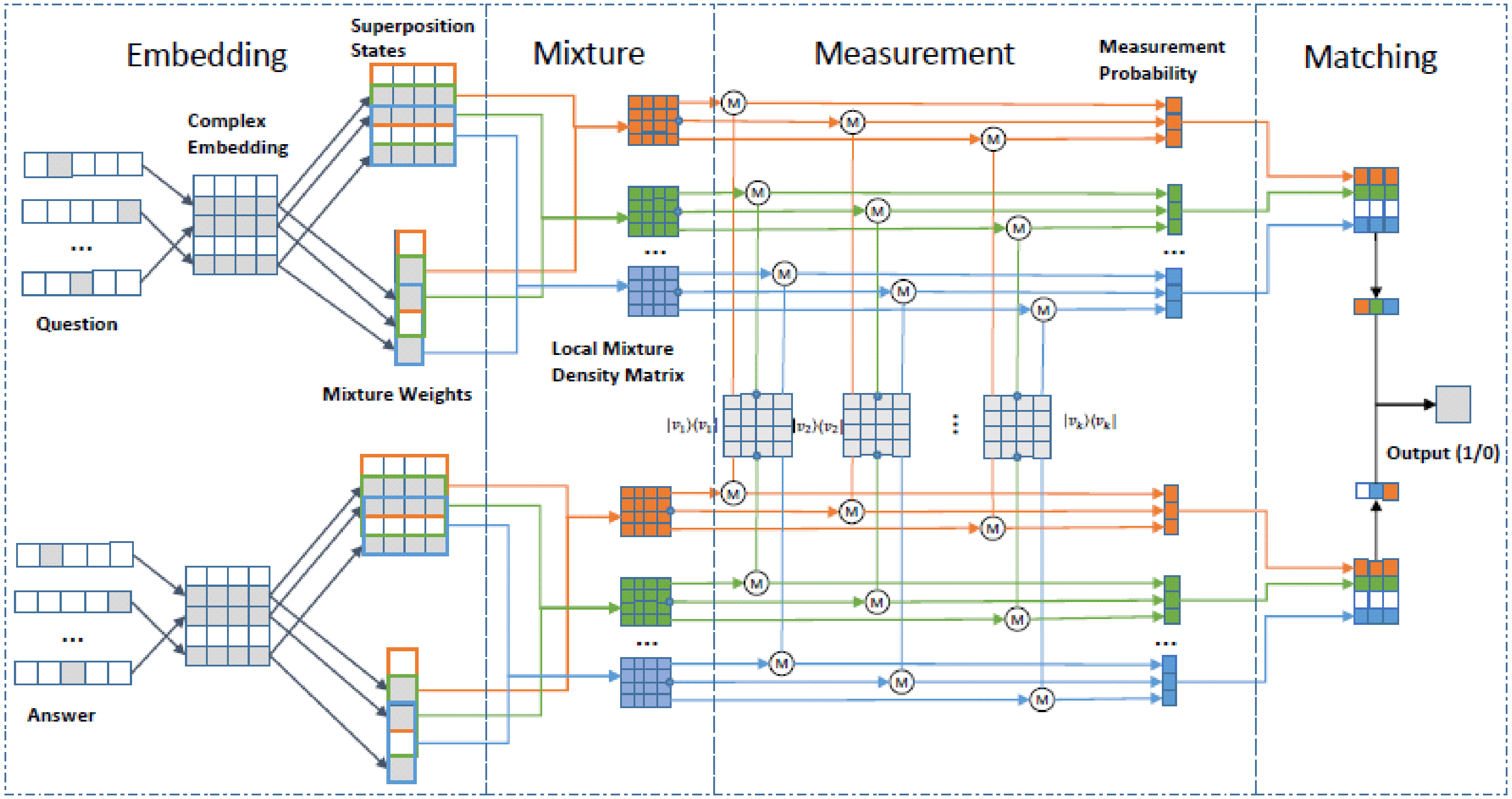
Case study for our measurement

Table 7: 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

Implements for matching

Figure 1: Architecture of Complex-valued Network for Matching. Ⓜ means a measurement operation according to Eq. 2.



Case study

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

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 ...

Experiments

Table 3: Experiment Results on TREC QA Dataset. The best performed values are in bold.

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 4: Experiment Results on Yahoo QA Dataset. The best performed values are in bold.

Model	P@1	MRR
Okapi BM-25	0.2250	0.4927
LSTM	0.4875	0.6829
CNN	0.4125	0.6323
CNTN	0.4654	0.6687
QLM	0.3950	0.6040
NNQLM-I	0.4290	0.6340
NNQLM-II	0.4660	0.6730
CNM	**0.4880**	**0.6845**
Over NNQLM-II	4.72% ↑	1.45% ↑

Table 5: Experiment Results on WikiQA Dataset.The best performed values for each dataset are in bold.

Model	MAP	MRR
Bigram-CNN	0.6190	0.6281
BILSTM	0.6557	0.6695
LSTM-attn	0.6639	**0.6828**
CNN	**0.6701**	0.6822
QLM	0.5120	0.5150
NNQLM-I	0.5462	0.5574
NNQLM-II	0.6496	0.6594
CNM	0.6548	0.6664
Over NNQLM-II	1.01% ↑	1.01% ↑

Weights

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

	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

Learned measurements

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

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

Ablation Test

Table 9: Ablation Test. The values in parenthesis are the performance difference between the model and CNM.

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

Conclusion

- More concrete physical meaning
- Self-explainable subcomponents
- More constrain for the subcomponents
- Guided by Quantum probability theory

Future works with this topic

- Explore high-dimension **tensor network** with Quantum representation
- **Capsule** Network with Quantum insights
- Incorporating more knowledge (e.g. word Polarity) in phase part
- Multi-task setting to transfer learned measurement to similar tasks
- Exploring the language generating task (unitary transform)
- Cross-language entanglement
- Exploring position-aware quantum representation for image