DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

H2020-MSCA-ITN Grant Agreement N. 721321

QUARTZ
Quantum Information Access and Retrieval Theory

Dynamic Content Monitoring and Exploration using Vector Spaces (ESR-2)

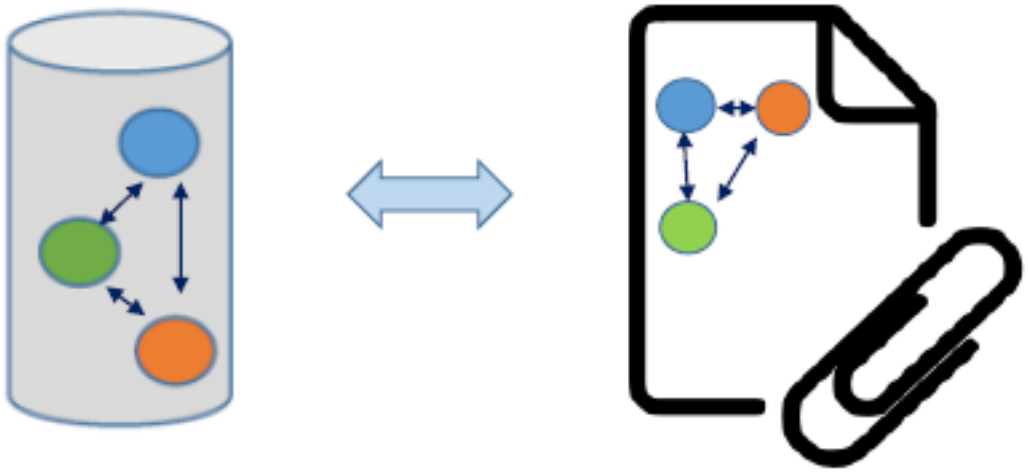
Benyou Wang

University of Padua

QUARTZ workshop, Copenhagen

03/04/2019

Vision



Representing and interpreting language inspired by Quantum Theory

State-of-the-art Paradigm for Language

How to compose a sequence of words

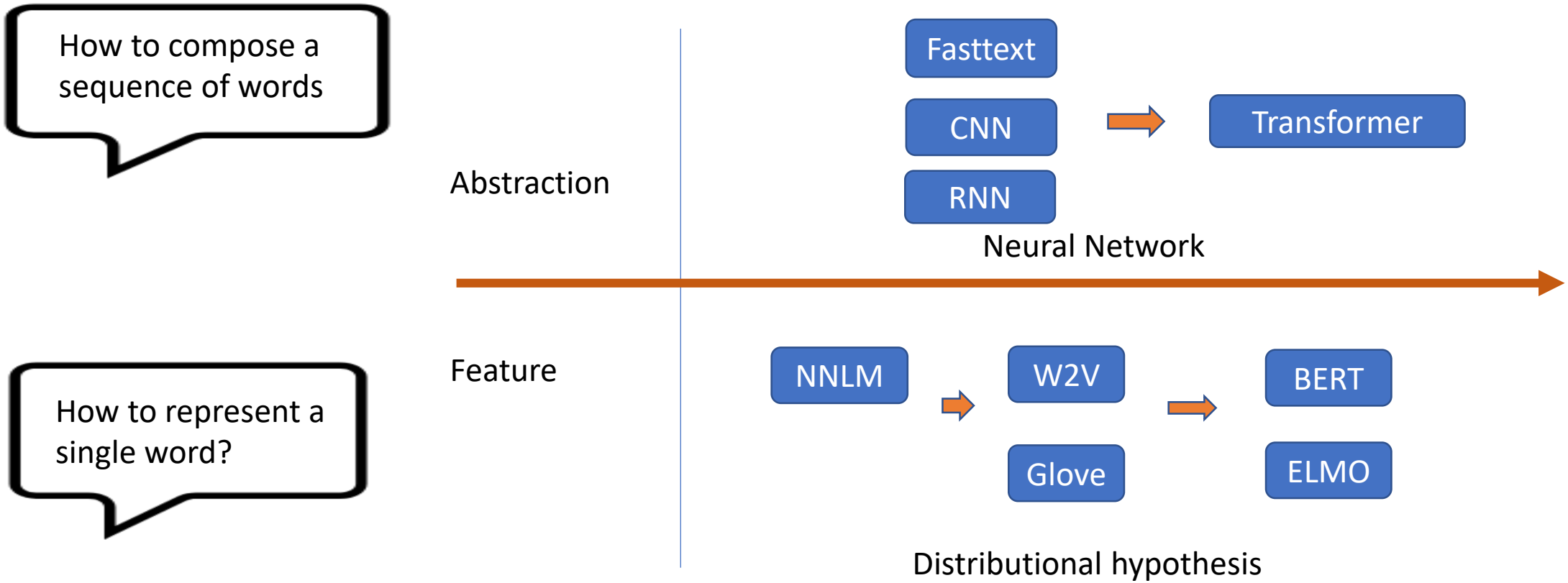
Abstraction

How to represent a single word?

Feature

Benyou Wang, Emanuele Di Buccio, Massimo Melucci. Representing Words in Vector Space and Beyond. Submitted to one Springer chapter organized by Andrei Khrennikov.

State-of-the-art Paradigm for Language



Benyou Wang, Emanuele Di Buccio, Massimo Melucci. Representing Words in Vector Space and Beyond. Submitted to one Springer volume edited by Aerts, Khrennikov and Melucci.

Concerns (1) - Representation

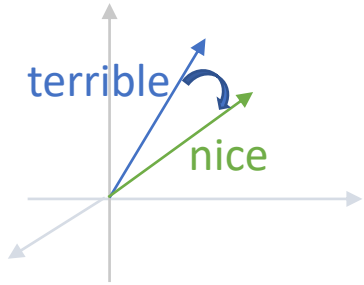
How to compose a sequence of words

How to represent a **single word**?

Abstraction

Feature

The weather is terrible in Copenhagen
The weather is nice in Copenhagen



Limitation of Distributional hypothesis

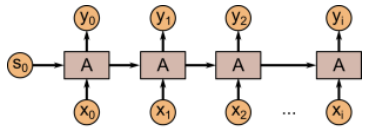
Concerns (1) - Representation

How to compose a **sequence** of words

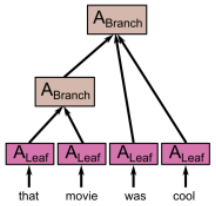
How to represent a single word?

Abstraction

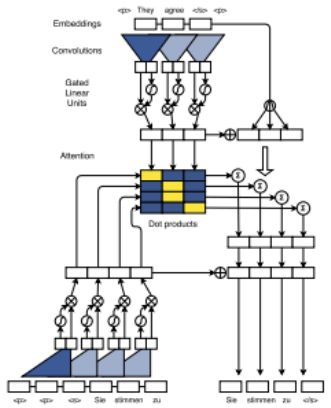
Feature



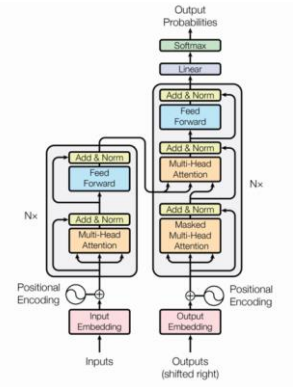
Recurrent Neural Net



Recursive Neural Net



Conv seq2seq



Transformer

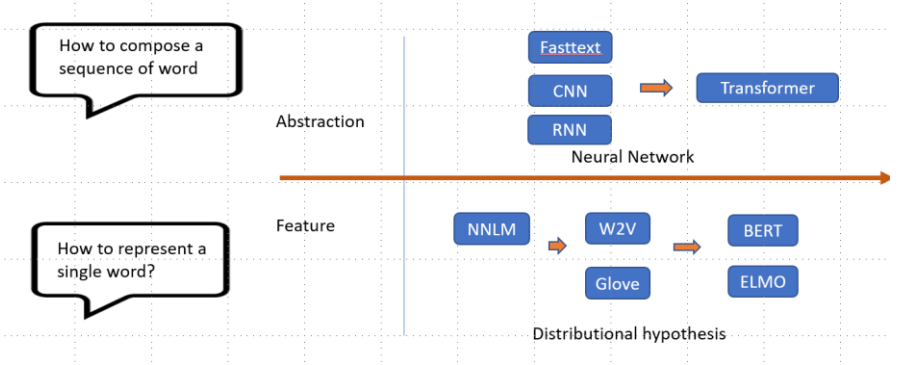
How to model sequential tokens?

https://pseudoprofound.wordpress.com/2016/06/20/recursive-not-recurrent-neural-nets-in-tensorflow/

Gehring, Jonas, et al. "Convolutional sequence to sequence learning." *ICML* 2017.
Vaswani, Ashish, et al. "Attention is all you need." NIPS 2017.

Concerns (2) - Interpretation

Kids store **1.5 megabytes** of information to master their native language ?



So complicated !!!

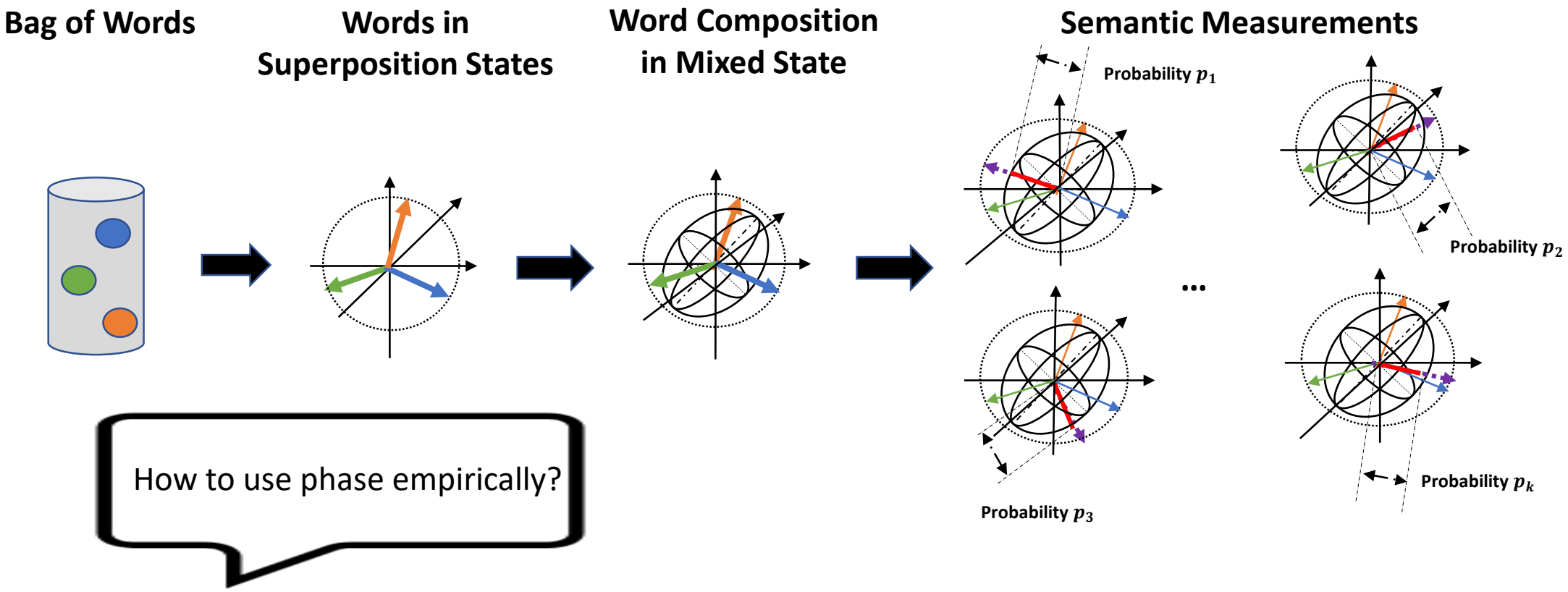
How to interpret/understand this empirical framework with a more robust way, like from a Mathematical/Physical perspective?

https://news.berkeley.edu/2019/03/27/younglanguagelearners/

Contents

- Representations: Towards **Complex-valued** word embedding
  - Limitation of the Distributional Hypothesis
    - **Sentiment-aware** Complex word embedding
  - Extending sequential abstraction
    - Encoding **position** in Complex word embedding
- Interpretation: From **higher-dimensional** Hilbert Space
  - Rethinking the neural network based NLP Paradigm in Tensor perspective

Semantic Hilbert Space



Benyou Wang*, Qiuchi Li*, Massimo Melucci, Dawei Song. Semantic Hilbert Space for Text Representation Learning. **WWW 2019**
Qiuchi Li*, **Benyou Wang***, Massimo Melucci. CNM, a Complex-valued Matching Network for Matching. NAACL **2019**

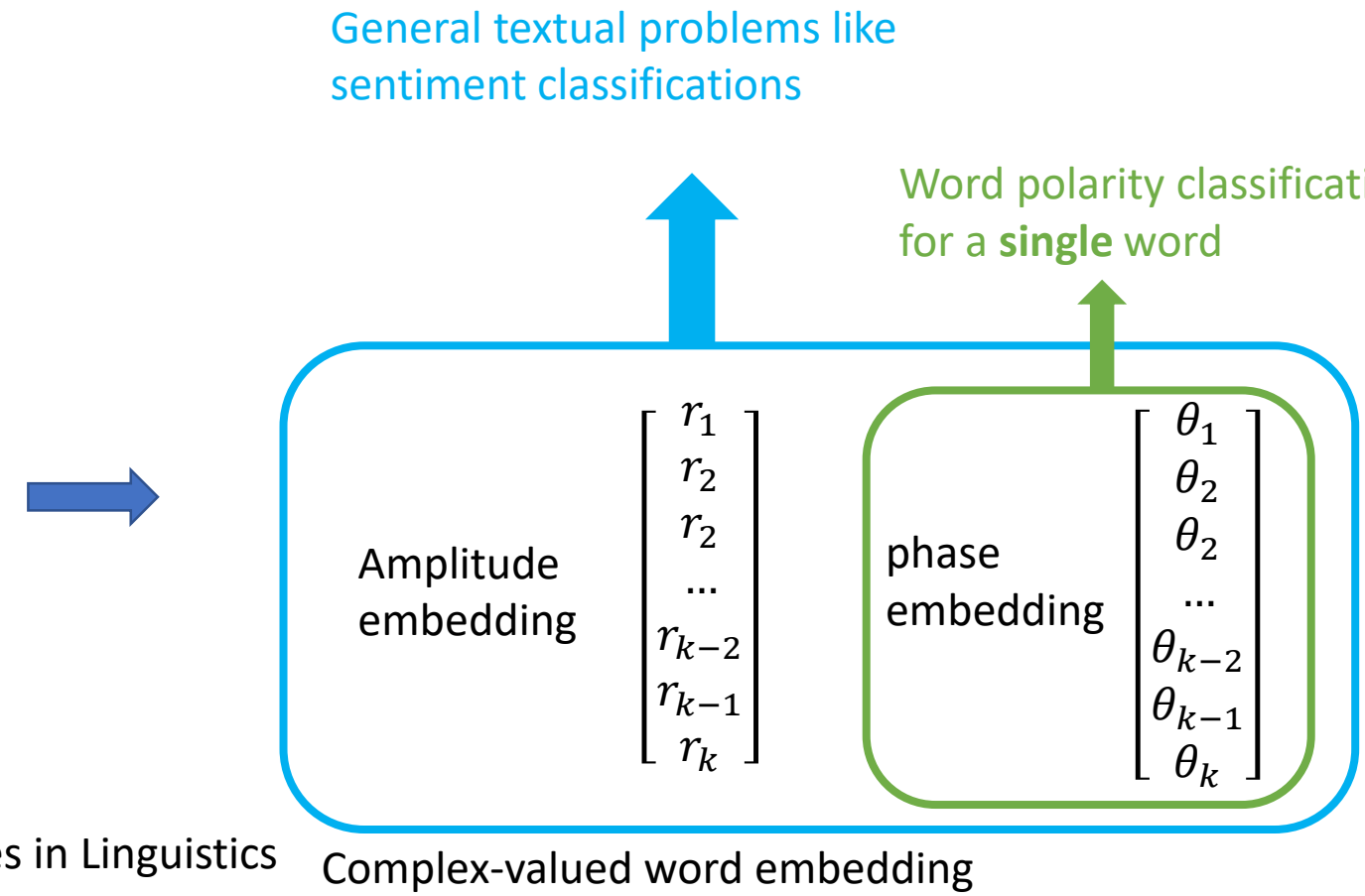
Sentiment-aware complex word embedding

$$|w\rangle = \sum_{j=0}^{k} r_j e^{-i\theta_i} |e_j\rangle$$

$\{|e_j\rangle\}_{j=0}^{100}$= is a one-hot vectors,

e.g., $e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \dots \\ 0 \\ 0 \\ 0 \end{bmatrix}$ and $e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \dots \\ 0 \\ 0 \\ 0 \end{bmatrix}$

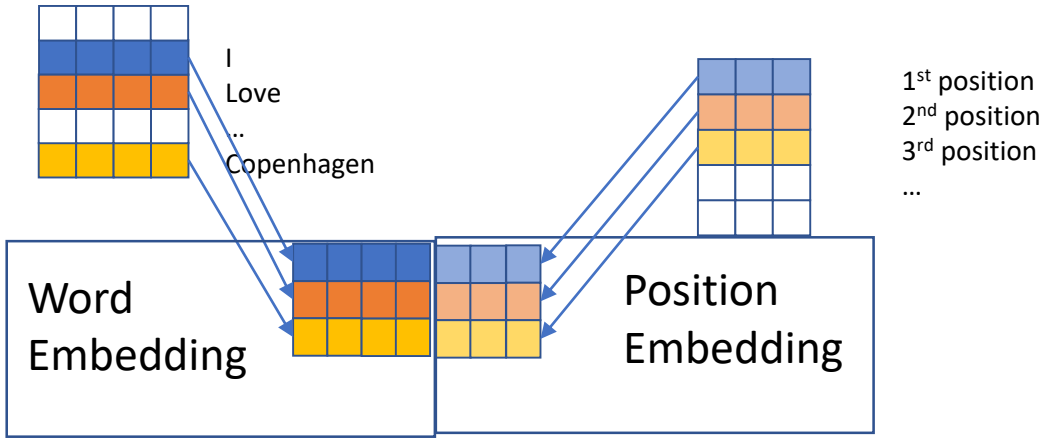
They can be considered as sememes in Linguistics



The idea was discussed with Qiuchi Li, Sagar Uprety and Chen Zhang.

Position-aware complex word embedding

Position embedding in Conv Seq2seq [1]



"I love Copenhagen"

Position embedding in Transformer[2]

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

Do not need training !!!

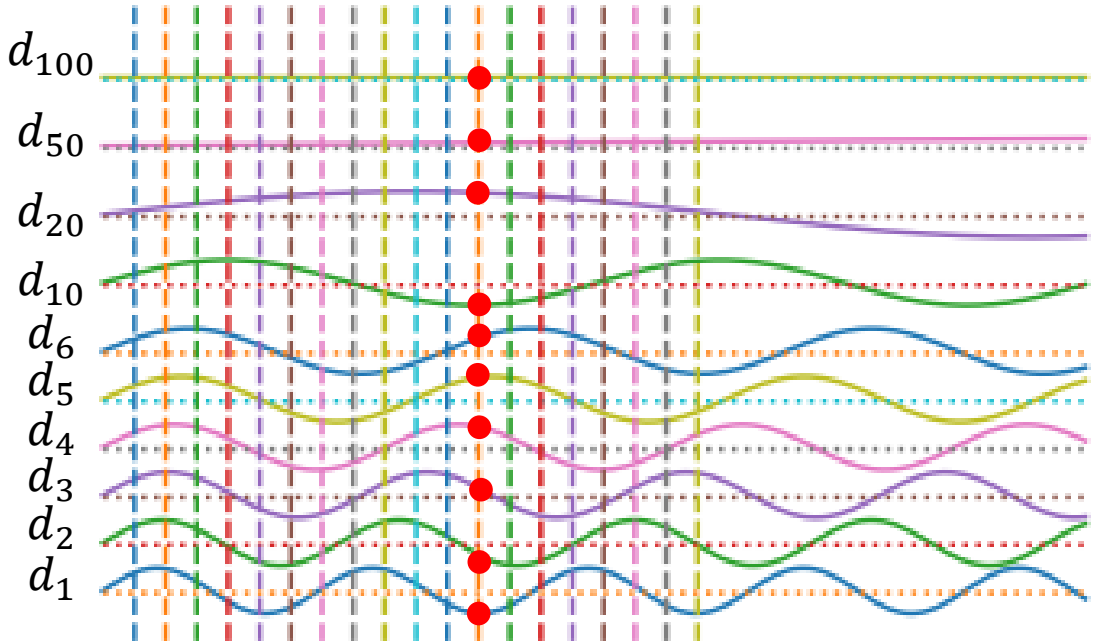
We proved that the position embedding in Transformer can be **derived from Complex-valued word embedding**

[1] Gehring, Jonas, et al. "Convolutional sequence to sequence learning." *ICML* 2017.
[2] Vaswani, Ashish, et al. "Attention is all you need." NIPS 2017.

The idea was discussed with Donghao Zhao and Qiuchi Li.

Position-aware complex word embedding

$$|w\rangle = \sum_{j=0}^{100} r_j e^{-i\theta_i} |e_j\rangle = \sum_{j=0}^{100} (r_j \cos\theta_j + i * r_j sin\theta_j) |e_j\rangle$$



$\cos\theta_j$

$Real(|w\rangle) =< \vec{\bullet} , \vec{r} >$

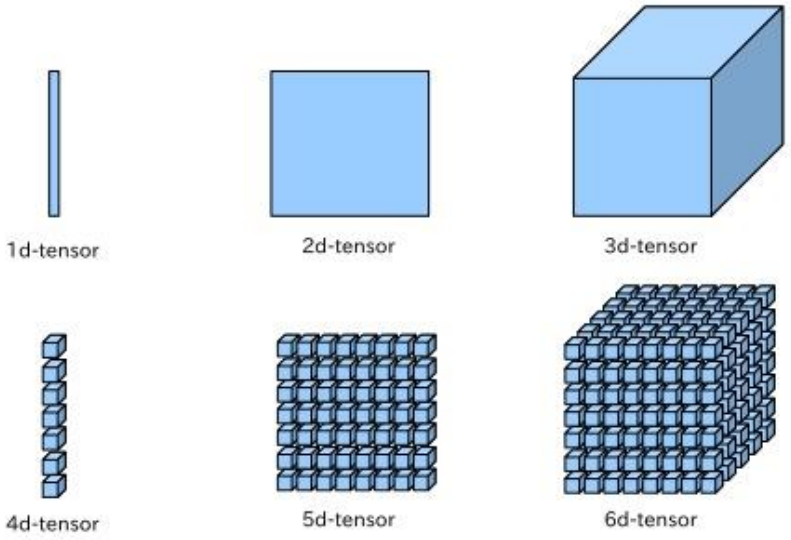
Word in different position has different phase, resulting in different word embedding

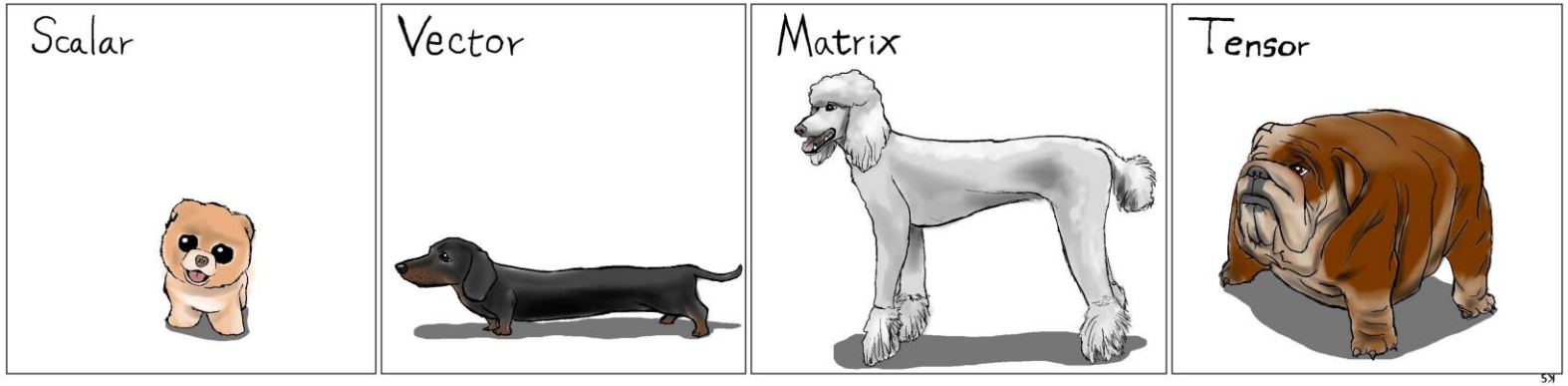
https://gist.github.com/wabyking/83ab87327e707b3e2834e2f37cac6bcf is used to draw the figure
The idea was discussed with Donghao Zhao and Qiuchi Li.

Contents

- Representations: Towards **Complex-valued** word embedding
  - Limitation of the Distributional Hypothesis
    - **Sensitive-aware** Complex word embedding
  - Extending sequential abstraction
    - Encoding **position** in Complex word embedding
- Interpretation: From **higher-dimensional** Hilbert Space
  - Rethinking the neural network based NLP Paradigm in tensor perspective

What is tensor?





The number of publications on tensor increases **exponentially** over recent years;

QUARTZ
Quantum Information Access and Retrieval Theory

Tensor is everywhere– find the third and higher dimensions

- **Modality** : Multi-modal Data
  - $[x_{image}, x_{text}, x_{sound}, \ldots]$ where x is the feature matrix or $H_{image} \otimes H_{text} \otimes H_{sound} \otimes \ \ldots$
- **Time** : Time-series Arxiv paper collections
  - $[D_{t1}, D_{t2}, \ldots D_T]$, where D is the word-word co-occurrence matrix
- **Domain**: Multiple-domain wiki pages
  - $[D_{nature}, D_{science}, \ldots, D_{engineering}]$, where D is the word-word co-occurrence matrix
- **Word sequence**: text
  - $H_I \otimes H_{Love} \otimes H_{Copenhagen} \otimes \cdots \otimes H_{\sim}$
- **Relation** : hyperlink data (anchor text as the relation)
  - $H_{source} \otimes H_{relation} \otimes H_{target}$
- **Cross-feature regression :** regression with x features
  - Hypotheses space for a general regression model:

$$H_y = b + \sum_{i \in N} b_i x_i + \sum_{i,j \in N} b_{i,j} x_{i,j} + \cdots + \sum_{i,j \in N} b_{i,j,\ldots k}\, x_{i,j,\ldots k}$$

Tensor representation in Text

For a sentence “I Love Copenhagen”

$$|\Phi_s\rangle = |w_1\rangle \otimes |w_2\rangle \otimes |w_3\rangle \otimes \cdots |W_n\rangle$$

e.g., $|I\rangle \otimes |love\rangle \otimes |Copenhagen\rangle$

One-hot Tensor representation

$$|w\rangle = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \dots \\ 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow |\Phi_s\rangle \in R^{|V|^N}$$

Distributed Tensor representation

$$|w\rangle = \sum_{j=0}^{k} r_j e^{-i\theta_i} |e_j\rangle \Rightarrow |\Phi_s\rangle \in R^{k^N}$$

- Generally speaking, $|V| \gg k$

Hypotheses space with tensor representation

Hypotheses space for a specific label y, with a sentence S

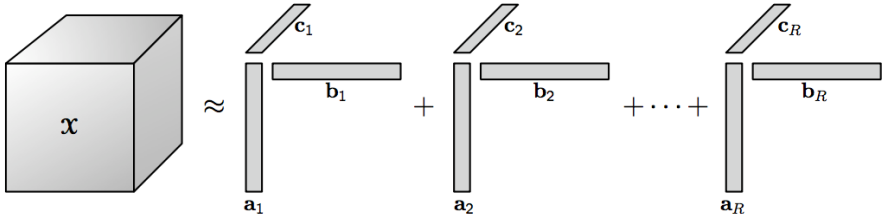
$$H_y(|\Phi_s\rangle) = < W, |\Phi_s\rangle >$$

Where $<\cdot,\cdot>$ is inner product, which denotes a sum of element-wise multiplication.
W is a tensor, which has the dimension with $|\Phi_s\rangle$, including weights/coefficient for features $|\Phi_s\rangle$

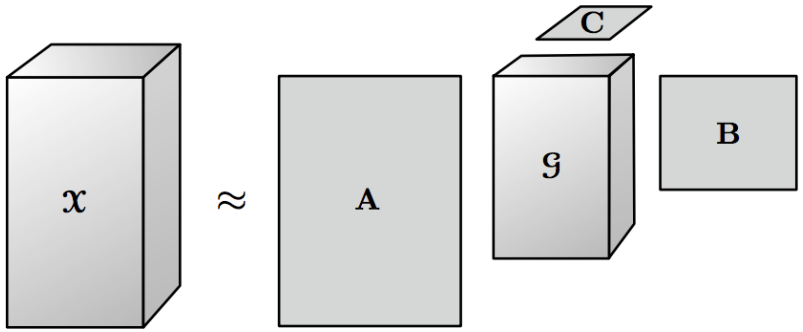
Attention !!! W is $R^{|V|^N}$ in one-hot word tensor representation
or R^{k^N} for distributed word tensor representation
We need to decompose it, in order to calculate $H_y(|\Phi_s\rangle)$

Tensor Decomposition

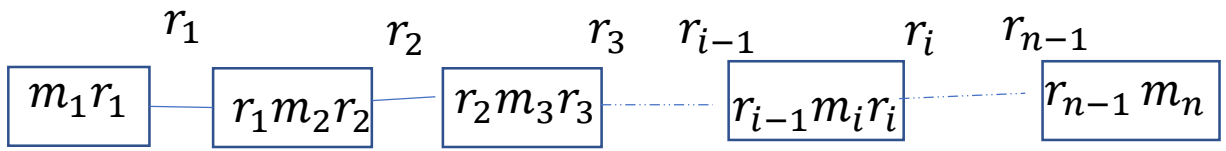
CP Decomposition $A = \sum_{z=1}^{Z} v_z^{(1)} \otimes \cdots \otimes v_z^{(N)}$



Tucker Decomposition



Tensor train (TT) Decomposition



$A = \mathcal{G}_1 * \mathcal{G}_2 * \ldots \mathcal{G}_n \in \mathcal{R}^{m_1 \times m_2 \times \ldots \times m_n}$

$\mathcal{G}_1 \in \mathcal{R}^{m_1 \times r_1}$

$\mathcal{G}_2 \in \mathcal{R}^{r_1 \times m_2 \times r_2}$

...

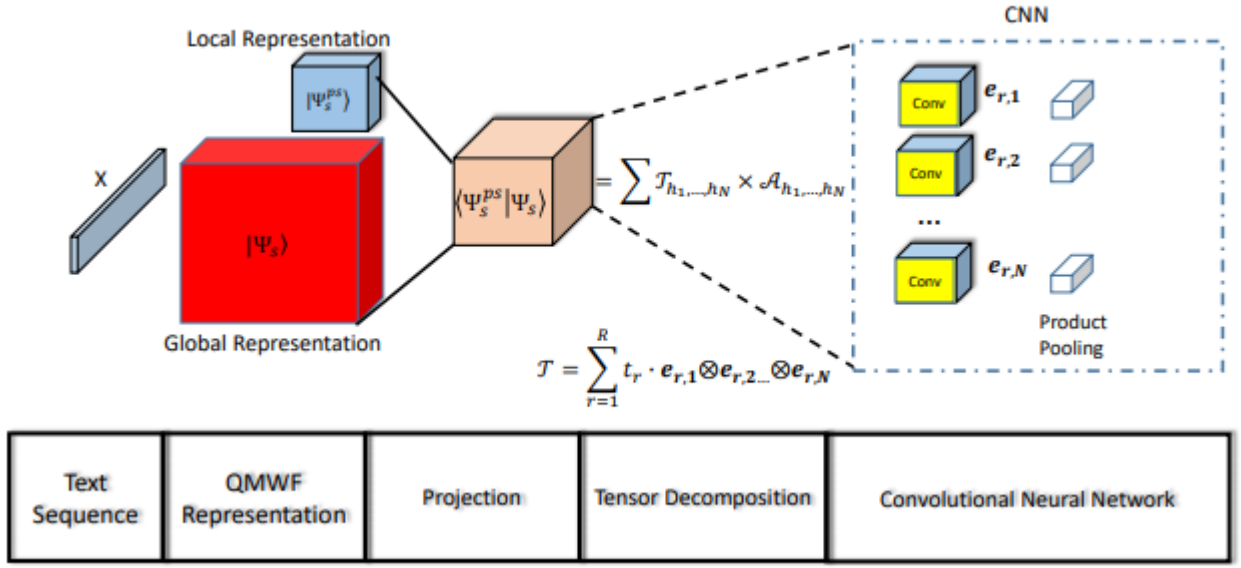
$\mathcal{G}_i \in \mathcal{R}^{r_{i-1} \times m_i \times r_i}$

...

$\mathcal{G}_n \in \mathcal{R}^{r_{n-1} \times m_n}$

Tensors are decomposed to a **sequential multiplication of some 3-order tensors**

CP Decomposition for distributed text tensor

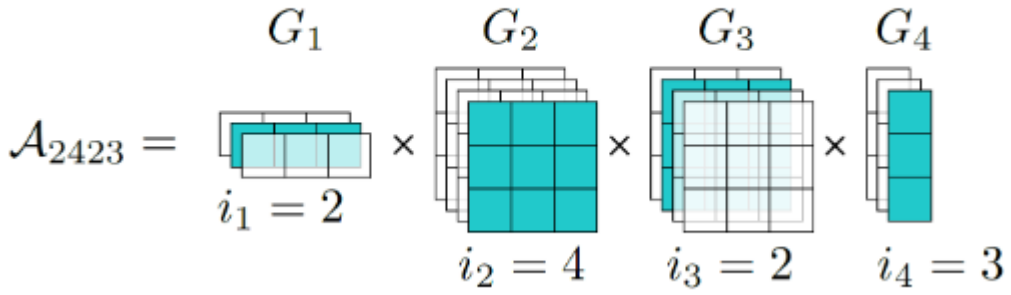


Zhang P, Su Z, Zhang L, **Wang B**, Song D. A quantum many-body wave function inspired language modeling approach. In Proceedings of the 27th ACM **CIKM** 2018 Oct 17 (pp. 1303-1312). ACM.

TT-Decomposition for one-hot text tensor

One-hot tensor

$$H_y(|\Phi_s\rangle) = < W, |\Phi_s\rangle >$$



A computing demo of $H_y(|\Phi_s\rangle)$, if $|\Phi_s\rangle$ is one-hot

It can be rewritten as

$$H_y(|\Phi_s\rangle) = < W, |\Phi_s\rangle = v_{start} M_{w_1}^{(1)} * M_{w_1}^{(2)} * M_{w_1}^{(3)} * \ldots M_{w_1}^{(n)} \; v_{end}$$

After sharing the weights in different positions, we have

$$H_y(|\Phi_s\rangle) = < W, |\Phi_s\rangle = v_{start} M_{w_1} * M_{w_1} * M_{w_1} * \ldots M_{w_1} \; v_{end}$$

Which admits a map $f : N \rightarrow R^{r*r}$, i.e., map a word index to a matrix

Research Activities

- Recent activities
  - Courses in University of Padova, like Applied Functional Analysis.
  - English course including speaking, grammar and listening.
  - Publications:
    1. Qiuchi Li*, **Benyou Wang***, Massimo Melucci. A Complex-valued Network for Matching. **NAACL 2019 (oral presentation)**
    2. **Benyou Wang***, Qiuchi Li*, Massimo Melucci, Dawei Song. Semantic Hilbert Space for Text Representation Learning. **WWW 2019 (short paper)**
    3. Wei Zhao*, **Benyou Wang***, Min Yang, Jianbo Ye, Zhou Zhao, Xiaojun Chen, Ying Shen. Leveraging Long and Short-term Information in Content-aware Movie Recommendation via Adversarial Training. **IEEE Transactions on Cybernetics (TOC), 2019**

- The following research activities:
  - Conferences : I will attend the **WWW 2019** and **NAACL 2019** if visa can be ready
  - August – October 2019 : Secondment in University of Copenhagen
  - October – December 2019 : Secondment in University of Montreal

* means equal contribution