School of CIT –
Social Computing
Research Group

Technical
University
of Munich

Advanced Natural Language Processing
CIT4230002

Prof. Dr. Georg Groh
M.Sc. M.Sc. Fabienne Marco

Lecture 7.2
Causality Applied

Lecture Outline – Lecture 1

- Causal Reinforcement Learning
- Causal BERT
- Compositionality of Language
- Quantum Natural Language Processing

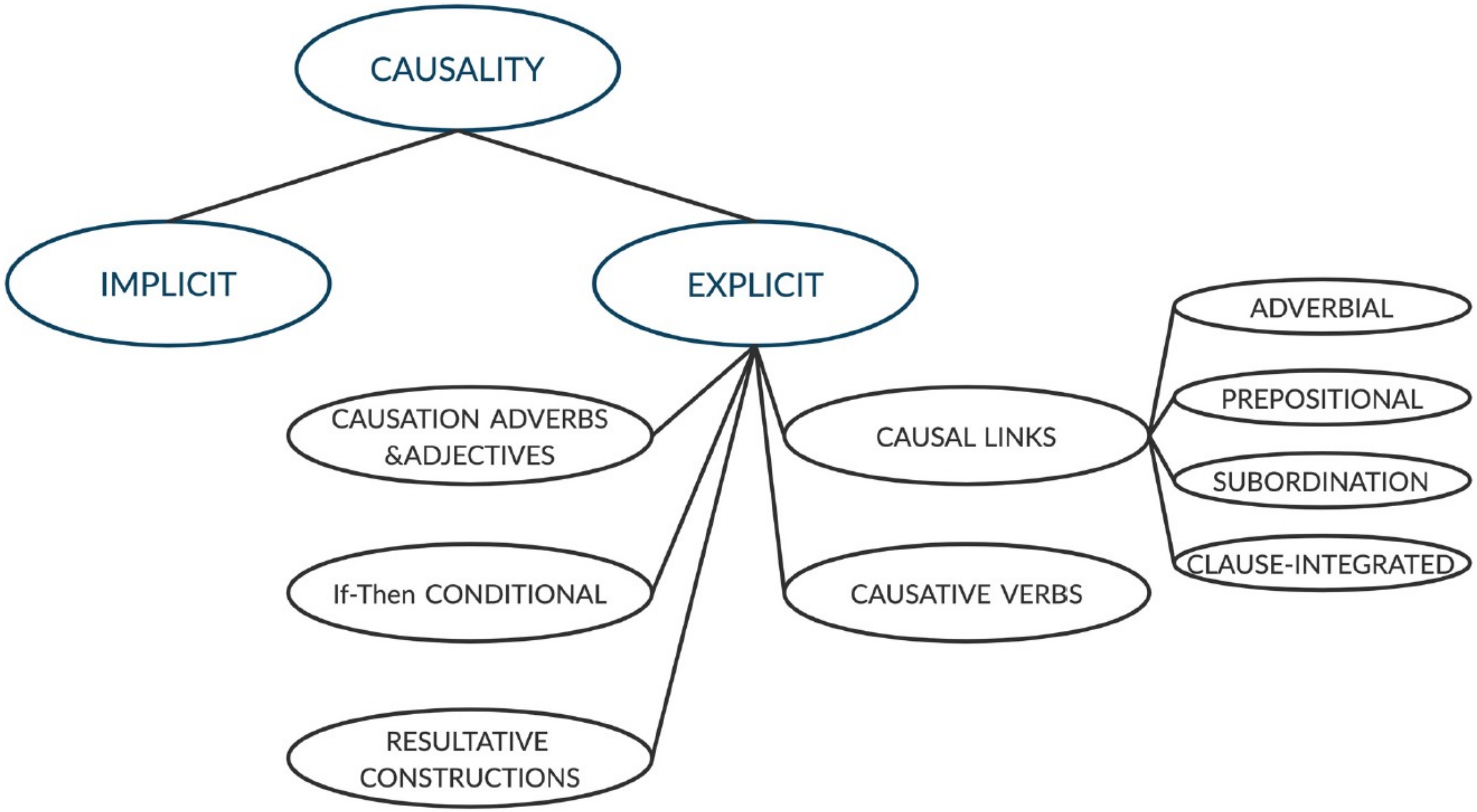


Fig. 1. Causality in natural language

Lecture Outline – Lecture 2

- Causal Reinforcement Learning
- CausalBERT
- Compositionality of Language
- Quantum Natural Language Processing

Causal Reinforcement Learning

Historical Context:

- Up to the 18th century, scientists sought causal explanations in accessible models like Linear Regression.
- The distinction between correlation and causation was later recognized.

Modern Machine Learning:

- Optimizes parameters based on training data, tracking correlations without causation.
- Structural causal models (SCMs) require domain knowledge and can't learn online.

Causal Reinforcement Learning (CRL):

- Combines Reinforcement Learning (RL) and Causal Reasoning.
- Aims to balance the advantages and disadvantages of both types of models.
- Examines CRL's basic idea, benefits, challenges, and future research potential.

Causal Reinforcement Learning

Foundations: CRL integrates RL and causal reasoning. Judea Pearl and Elias Bareinboim are pioneers in this field. Pearl's "The Book of Why" emphasizes causality's importance in developing general AI.

Reinforcement Learning (RL): Goal-oriented, adaptive learning through an agent's interaction with the environment. RL focuses on maximizing expected rewards but lacks causal assumptions.

Structural Causal Models (SCMs): Use directed acyclic graphs to model environments. Include graphical models, structural equations, and counterfactual/interventional logic.

Causal Reinforcement Learning

Concept: CRL incorporates causal graphs and causal hierarchy into RL. Models the agent's environment causally, enabling behavior transfer across different environments.

Challenges in CRL:

- Generalized Policy Learning: Combines online and offline strategies to deal with non-identifiability.
- When and Where to Intervene: Uses do-calculus to determine the impact of interventions.
- Counterfactual Decision-Making: Optimizes for regret instead of reward, allowing exploration of alternative scenarios.

Causal Reinforcement Learning

Transfer Learning: Zhang and Bareinboim's work on transfer learning from systems like Multi-Armed Bandits using CRL. Focuses on identifying causal effects and transferring learning across tasks.

Medical Applications: Zhang and Bareinboim's framework for dynamic treatment regimes in medical treatments. Uses causal graphs to optimize treatment strategies, particularly for chronic diseases.

Fairness in Decision-Making: Bareinboim and Pearl's research on detecting and addressing discrimination using SCMs. Differentiates between direct, indirect, and spurious discrimination in models.

Lecture Outline – Lecture 2

- Graph-Based Causal Models
- Causal Reinforcement Learning
- CausalBERT
- Compositionality of Language
- Quantum Natural Language Processing

CausalBERT

CausaLM (CausalBERT) is a language model specifically designed to understand and work with causal relationships in textual data. Below is a summary of what you need to know about CausalBERT

CausalBERT

- The different causal BERT models evolve one from to another

→Bert is used to deliver the context for Event aware CausalBert while Masked Causal C-BERT is using Event aware Causal Bert to finetune after masking the whole event

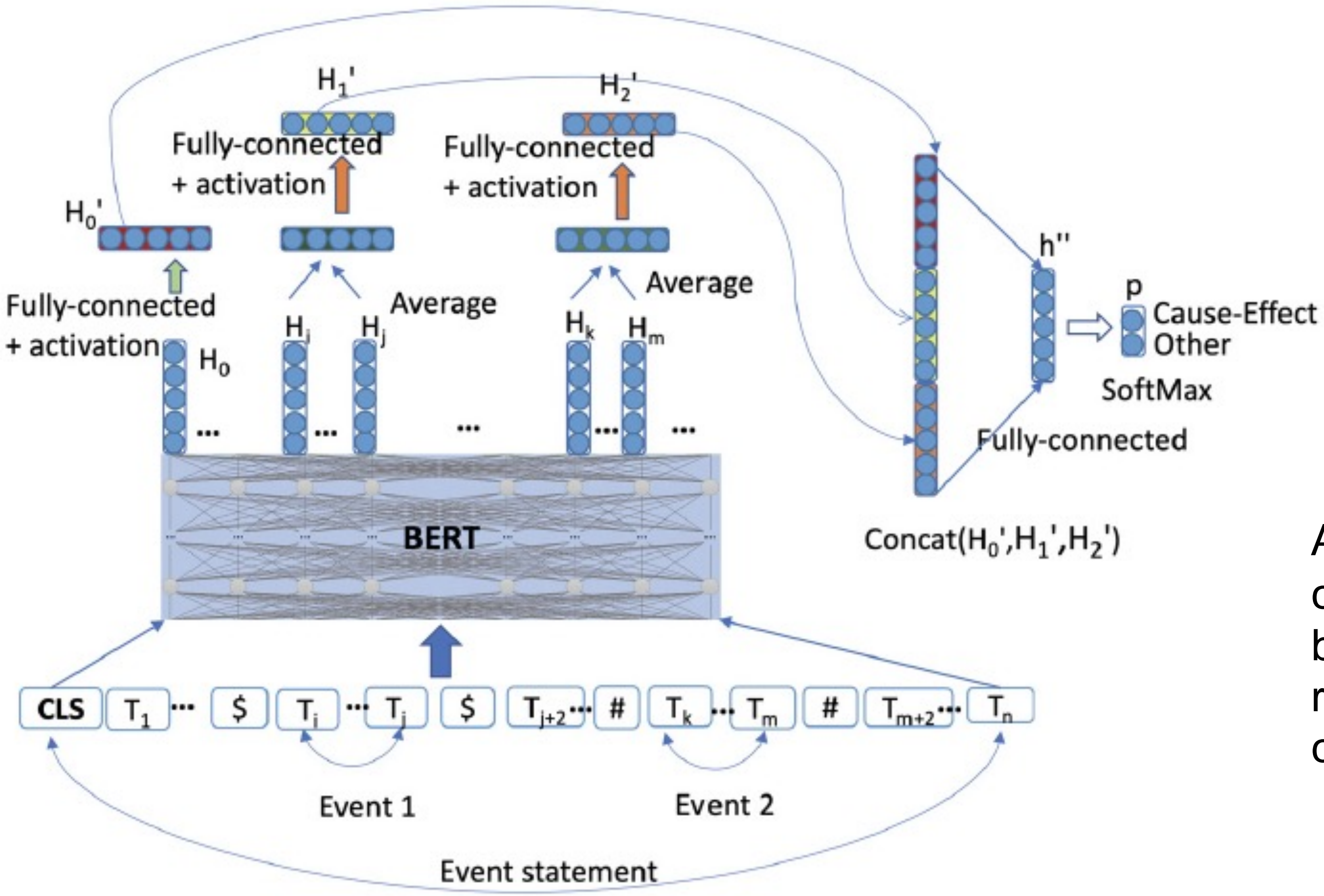


Fig. 3. Event Aware C-BERT

Averages the context of causal BERT to get a better context and reduce the amount of context vectors.

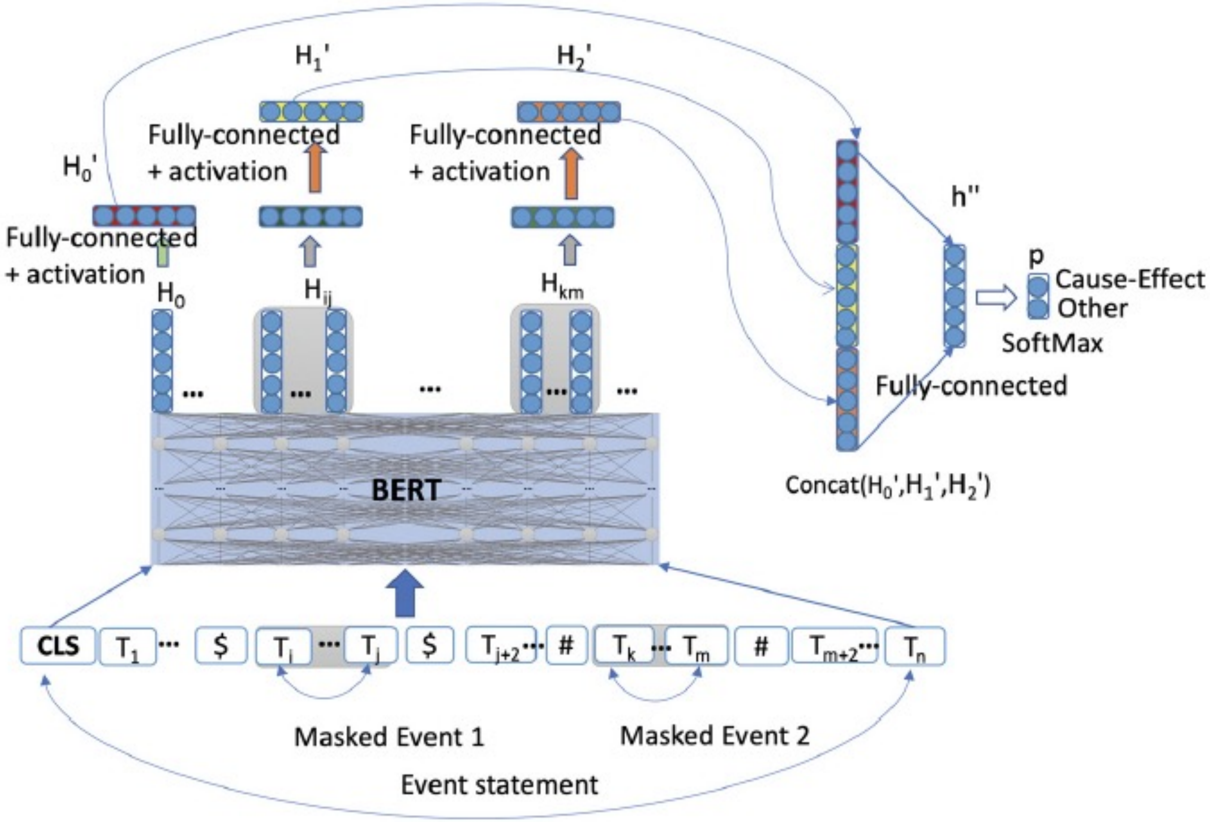


Fig. 4. Masked Event C-BERT

Includes the event-aware BERT for delivering context

CausalBERT evolution levels – let‘s look into it

Table 2. Example sentences, sentence with event markers, and masked event markers for curated datasets

Curated corpus			
Dataset	Example sentence	Sentence with event marker	Sentence with masked event marker
Semeval 2007	Most of the taste of strong onions comes from the smell	Most of the <**e1**> **taste** </**e1**> of strong onions comes from the <**e2**> **smell**</**e2**>	Most of the <**e1**> **blank** </**e1**> of strong onions comes from the <**e2**> **blank** </**e2**>.
Semeval 2010	As in the popular movie “Deep Impact”, the action of the Perseid meteor shower is caused by a comet, in this case periodic comet Swift-Tuttle	As in the popular movie “Deep Impact”, the action of the Perseid <**e1**> **meteor shower** </**e1**> is caused by a <**e2**> **comet** </**e2**>, in this case periodic comet Swift-Tuttle	As in the popular movie “Deep Impact”, the action of the Perseid <**e1**> **blank** </**e1**> is caused by a <**e2**> **blank** </**e2**>, in this case periodic comet Swift-Tuttle.
ADE	Quinine induced coagulopathy –a near fatal experience	<**e2**> **Quinine** </**e2**> induced <**e1**> **coagulopathy** </**e1**>–a near fatal experience	<**e2**> **blank** </**e2**> induced <**e1**> **blank** </**e1**>–a near fatal experience

Lecture Outline – Lecture 2

- Graph-Based Causal Models
- Causal Reinforcement Learning
- CausalBERT
- Compositionality of Language
- Quantum Natural Language Processing

Compositionality of Language l Definition

Compositional language is the principle that the meaning of a whole expression is a function of its parts and their syntactic arrangement. Historically, this idea was significantly shaped by Gottlob Frege in the late 19th and early 20th centuries, further developed by philosophers and linguists through the 20th century. Understanding compositionality is crucial for analyzing language structure, semantics, and the generative nature of human communication.

Productivity:

•Compositionality allows for the creation of an infinite number of meaningful sentences from a finite set of elements (words and grammatical rules).

Predictability:

•Given the meanings of individual words and the rules for combining them, one can predict the meaning of the entire expression.

Generativity:

•New sentences and meanings can be generated, allowing speakers to express novel ideas and thoughts

Compositionality of Language l Example Argument

Evolutionary Origins of Compositionality

•**Complexity:** The evolutionary origins of compositional language and its development in human history.

•**Details:** Compositionality likely provided significant adaptive advantages by enabling complex communication and abstract thought. The transition from non-compositional to compositional communication systems involves significant cognitive and social shifts.

•**Implication:** Exploring the evolutionary roots of compositionality can reveal how language may have co-evolved with other cognitive abilities, such as theory of mind and social cognition. It also helps in understanding the uniqueness of human language compared to other animal communication systems.

Compositionality of Language| Mathematical Foundations

- Quantum Systems and Algorithms operate on a mapping in the complex space
- Therefore, quantum systems operate on (complex) Hilbert spaces
- The n-dimensional tensor product serves as standard scalar product for quantum modules.
- Categorical theory is used to describe the different grammatical structures used within QNLP
    1. Categories
    2. Functors
    3. Natural Transformations
    4. Universality
    5. Adjoints
- Category: Verbs, abstract grammatical obejcts, arguments
- Functor: Homomorphism between two categories + Homomorphism between the Homomorphisms within categories
- Adjoint: Join (concatenation) of a linear map in one direction

Lecture Outline – Lecture 2

- Graph-Based Causal Models
- Causal Reinforcement Learning
- CausalBERT
- Compositionality of Language
- Quantum Natural Language Processing

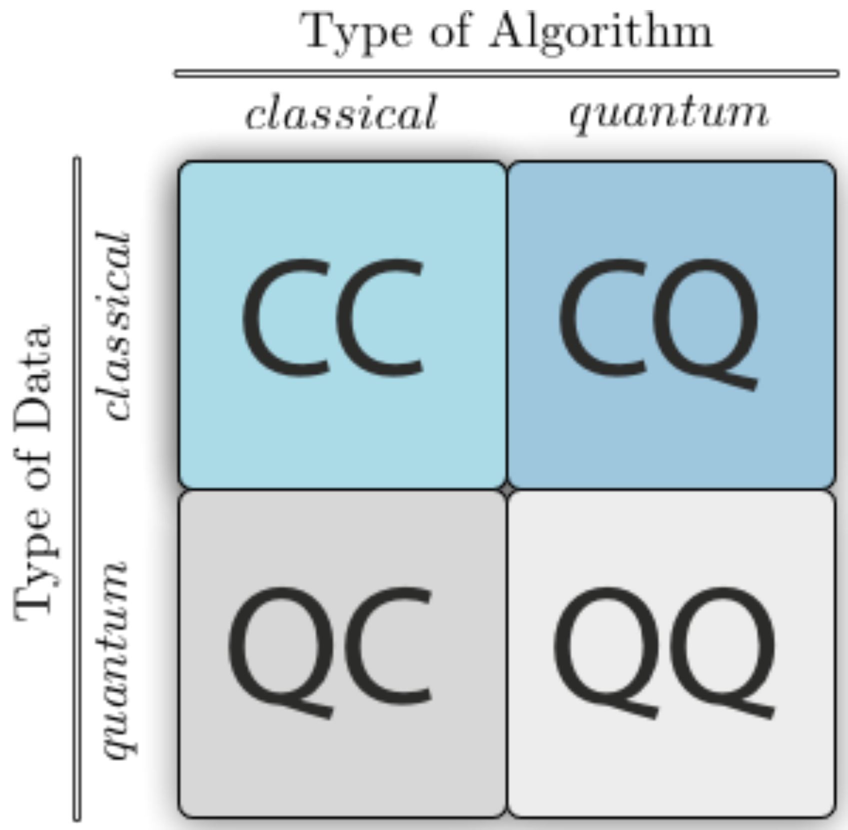
QNLP | Introduction

- Quantum Neural Networks trace back to the 90s

- Quantum Machine Learning and Quantum Neural Networks are used as simultaneous terms

- Quantum Machine Learning = Variational Quantum Circuits

- CQ = Advances are required
- QQ = Advantage is thought to be certain



Qubits is the basic unit of information in quantum computing. Similar to its classical counterpart, the bit, it can assume two distinct values of 0 or a 1. The difference is that whereas a bit must be either 0 or 1, a qubit can be 0, 1 or a superposition of both. Conventionally, possible states of a qubit are represented using the Dirac notation: $|0\rangle$ and $|1\rangle$.

$$|0\rangle := \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad and \quad |1\rangle := \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Definition 3.2.1 (Qubit System). A qubit system is mathematically represented by a unit vector of $\mathbb{C}$. The typical examples of states are $|0\rangle := \binom{1}{0}$ and $|1\rangle := \binom{0}{1}$. Different from a classical system, a qubit system has infinitely many states of arbitrary unit vectors in $\mathbb{C}^2$:

$$|\psi\rangle = (a, b)^T = a|0\rangle + b|1\rangle \tag{3.2}$$

with the normalisation condition $|a|^2 + |b|^2 = 1$

*please note the foundations given in this lecture are specifically chosen for the aim of this lecture and are no general or complete introduction to quantum basics

QNLP | Basics of Quantum*| Superposition

Classic Interpretation	Quantum Interpretation
Superposition describes when two quanti ties are added together to make another t hird quantity that is entirely different fro m the original two. In case we measure the state, we can me asure the third, resulting state directly	Quantum systems include small objects for w hich non-classical effects are observed. These objects can only have certain states. W ithin the superposition the state can be in an y linear combination of the finite states. In case we measure then the state is one of t he states before going onto superposition.

Superposition in a NLP sense

SuperFormer: Continual learning superposition method for text classification (traditional algorithm)

QNLP: In the context of NLP, the superposition gives the possibility to better manage some pervasive natural language phenomena, such as lexical ambiguity and polysemy

*please note the foundations given in this lecture are specifically chosen for the aim of this lecture and are no general or complete introduction to quantum basics

QNLP | Basics of Quantum* | Entanglement

$$(\alpha_0|0\rangle + \alpha_1|1\rangle)(\beta_0|0\rangle + \beta_1|1\rangle) = \alpha_0\beta_0|00\rangle + \alpha_0\beta_1|01\rangle + \alpha_1\beta_0|10\rangle + \alpha_1\beta_1|11\rangle$$

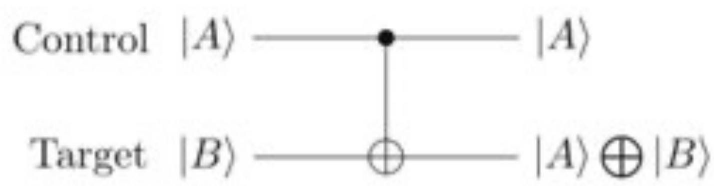
A product state can be represented in the fashion given above. Just like a classical system two separate probabilities are present for each measurement.

$$\frac{1}{\sqrt{2}}|00\rangle + \frac{1}{\sqrt{2}}|11\rangle \stackrel{?}{=} (\alpha_0|0\rangle + \alpha_1|1\rangle)(\beta_0|0\rangle + \beta_1|1\rangle),$$

$$\stackrel{?}{=} \alpha_0\beta_0|00\rangle + \alpha_0\beta_1|01\rangle + \alpha_1\beta_0|10\rangle + \alpha_1\beta_1|11$$

Before		After	
Control bit	Target bit	Control bit	Target bit
$\|0\rangle$	$\|0\rangle$	$\|0\rangle$	$\|0\rangle$
$\|0\rangle$	$\|1\rangle$	$\|0\rangle$	$\|1\rangle$
$\|1\rangle$	$\|0\rangle$	$\|1\rangle$	$\|1\rangle$
$\|1\rangle$	$\|1\rangle$	$\|1\rangle$	$\|0\rangle$

- What about such a state. Can it be separated?

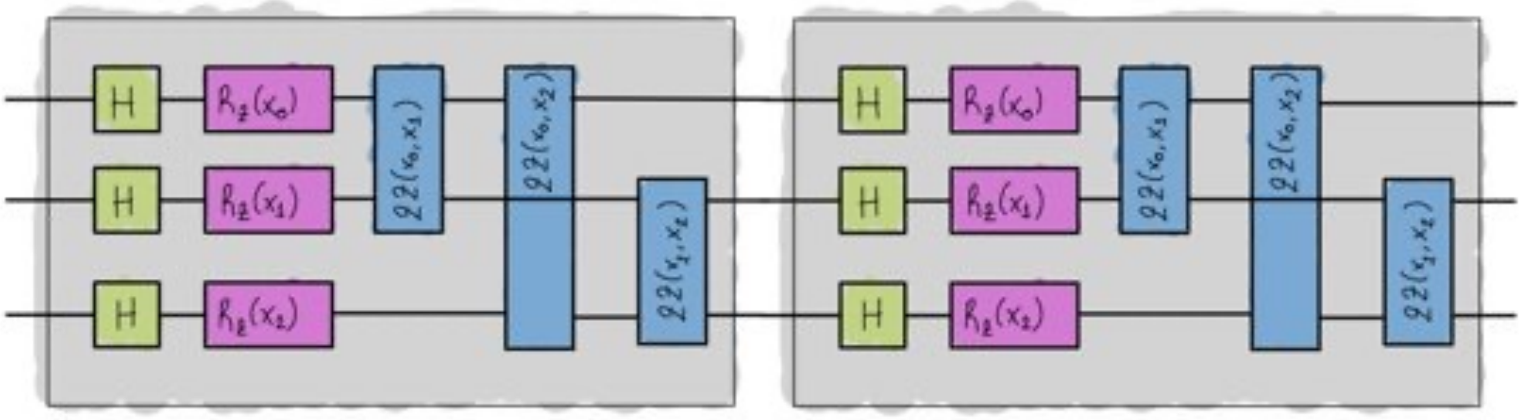


- This means that measurement of one of the qubits alter the probability associated with the measurement of the other. This is purely nonclassical and exactly what angered Einstein, Podolski and Rosen to publish(EPR paradox, spooky action at a distance) a piece arguing against quantum mechanics.

- They proposed there should be hidden variables underlying weirdness so
- that things can go back to being classically explainable.

*please note the foundations given in this lecture are specifically chosen for the aim of this lecture and are no general or complete introduction to quantum basics

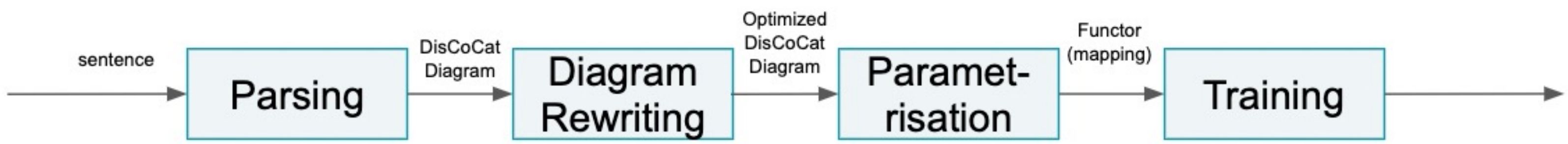
QNLP | Basics of Quantum*| Quantum Ansatz



- New ansatzes are continuously being developed and experimented with
- Most basic components of the design is parametrization and entangling
- Like their classical cousins quantum neural networks(QVAs) are also overparametrized
- This poses its advantages and disadvantages given problem types, data types
- Proclivity to get stuck on barren plateaus of 0-gradient
- Difficulty in calculation of gradients in the first place

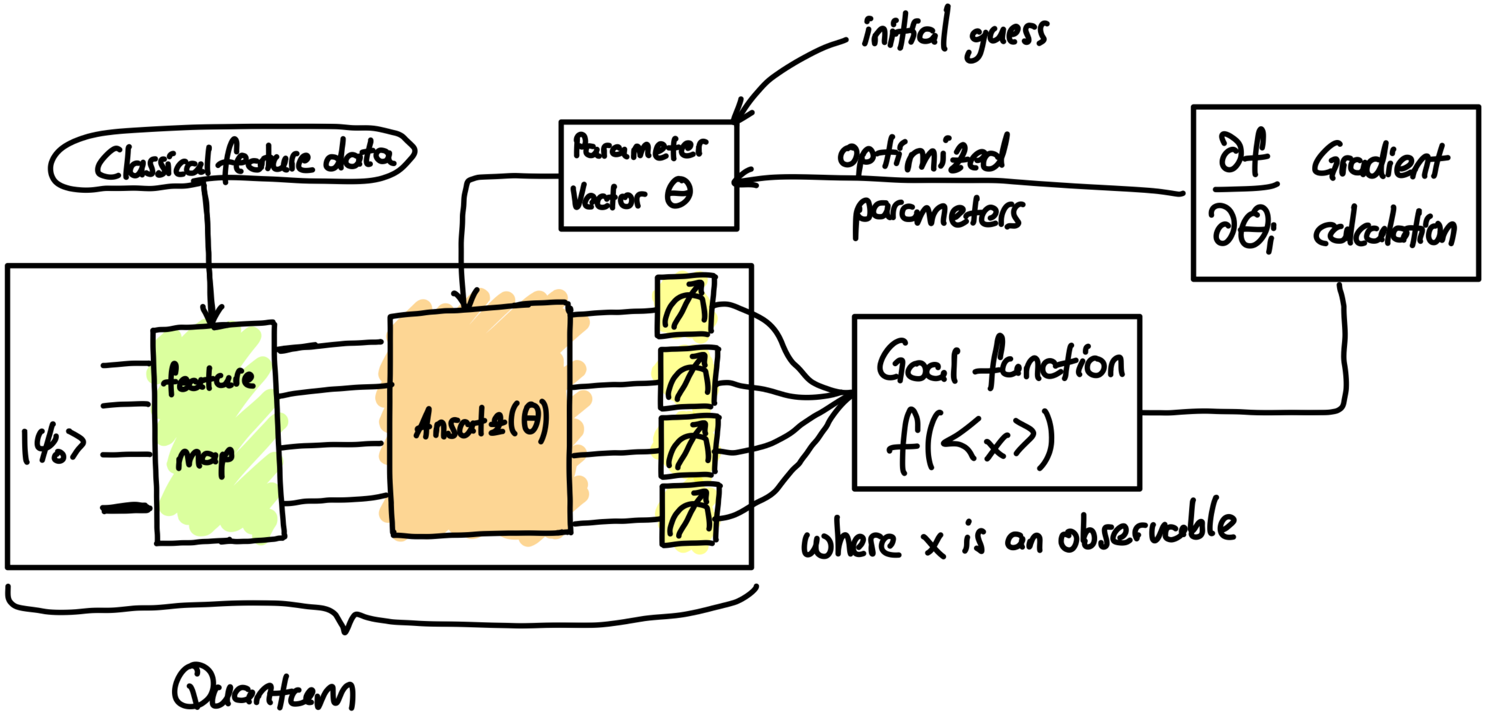
QNLP| QNLP Basics

- QNLP is based mainly on quantum simulation and/or hybrid models these days
- Most of the python frameworks introduced in this lecture can be run on a quantum computer directly – however most papers and we as well work with quantum simulation
- The Model is usually working the following way:



- The training process trains the received ansatz
- The parameters are not trained with repect to the loss function but with respect to the other gradients in the quantum circuit
- There is no closed form for the quantum optimisation process

$$\sigma^{t+1} = \sigma^{t} + \gamma \frac{\delta \mathcal{L}}{\delta \theta}|_{\theta=\theta^{t}}$$



- Through optimization(variation) of the parameters the circuit learns

QNLP | QNLP Basics | QNLP vs. NLP

	QNLP	NLP
Size	N-qubit system operates in the space of C^{2n}	**Dimension of the systems defined by the number of neurons (non-exponential)**
Architecture	Trainable Quantum Ansatz	**Various options like Encoder-Decoder Networks**
Training	**Parameters of the Variational Algorithm are trained;**	**Neural Network parameters are trained**
Cost Function	Cost is determined classically, though there are first ideas including it in the quantum circuits directly	Determination via cost function

QNLP | Tools for QNLP | DisCoCat

- DisCoCat (Distributional Compositional Categorical Framework)
- Based on the formalism of Neuman and Penrose‘s work on grammatical substitutes for tensor notations.
- Applications:
    - Word-sense ambiguity
    - Semantic Similarity
    - Question Answering
    - Machine Translation
    - Anaphora resolution

Advantages:

- Deals with ambiguity of language
- It is possible to tramsform arbitraty graphs based on category theory in quantum circuits (Tuomi, 2022)

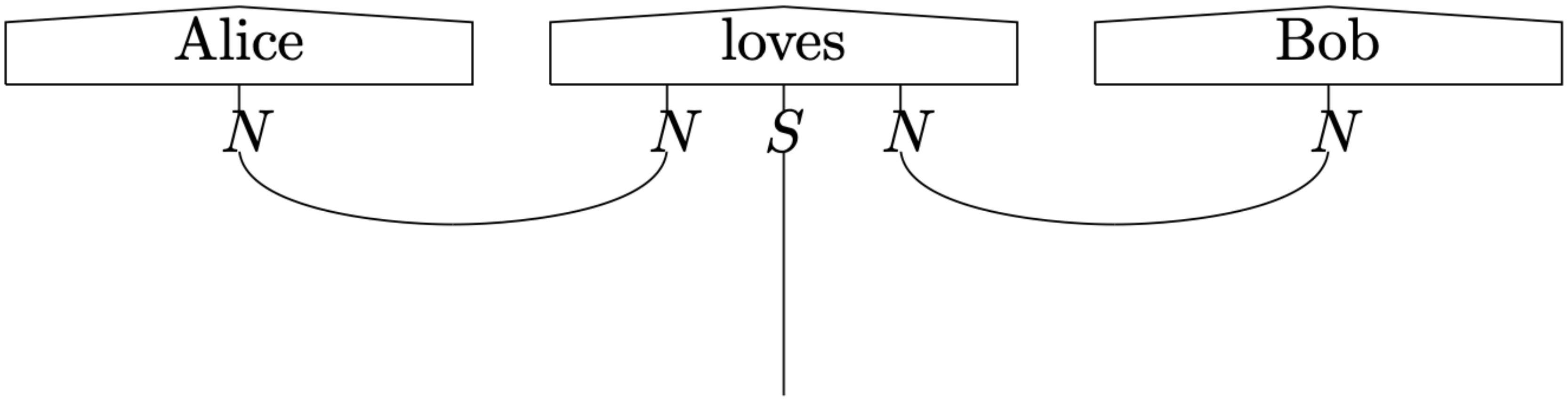
1. We translate a sentence by the help of a pregroup grammar

$$n^r \cdot s \cdot n^l$$

2. We reduce the sentence up to a single, measurable expression

$$n \cdot (n^r \cdot s \cdot n^l) \cdot n \leq 1 \cdot s \cdot 1 \leq s$$

3. This can be represented through different functors and then be turned into a parameterized quantum circuits



QNLP Models| Benchmarking by tasks

Task	Dataset	Models	Metrics	
			F1	Accuracy
Text classification	MELD dataset Sentiments (3-class)	CNN (Kim, 2014)	0.604	0.609
		RoBERTa (Liu et al., 2019)	0.721	-
		QIN (Zhang et al., 2019)	0.662	0.679
		QMN (Zhang et al., 2020)	0.729	0.756
	OMD dataset	Doc2vector (Le and Mikolov, 2014)	0.3979	0.6103
		SentiStrength (Thelwall et al., 2010)	0.6352	0.6110
		GQLM+QRE (Zhang et al., 2018c)	0.6261	0.6298
			Accuracy	
	SST-5	Star-Transformer (Guo et al., 2019)	53.0	
		BERT (Devlin et al., 2018)	52.9	
		BERT+TextTN (Zhang et al., 2021)	54.8	
			MAP	MRR
Question answering	WIKIQA	Bigram-CNN (Yu et al., 2014)	0.6190	0.6281
		AP-BILSTM (Santos et al., 2016)	0.6705	0.6842
		NNQLM-II (Zhang et al., 2018a)	0.6496	0.6594
		CNM (Li et al., 2019)	0.6748	0.6864
			MAP@10	NDCG@10
Information retrieval	TREC 2013	Unigram	4.91	6.05
		QLM (Sordoni et al., 2013)	6.14	6.70
		QLM-QE (Li et al., 2018)	8.94	10.37
			MAP	NDCG@20
	ClueWeb-09-Cat-B	MP (Pang et al., 2016)	0.066	0.158
		Conv-KNRM (Dai et al., 2018)	0.121	0.285
		QLM (Sordoni et al., 2013)	0.082	0.164
		QINM (Jiang et al., 2020)	0.134	0.338

Final Takeaways

- Causal Reinforcement Learning is quite helpful for analyzing specifically medical data, but it lacks suitable implementations for language
- BERT variants can deliver context and efficiently find cause-effect relations (downside: Just bicategorical classification)
- Language Compositionality is very explanatory but hard to implement
- QuantumNLP mainly builds on superposition and entanglement to mirror language ambiguity and resolves the main conflict by using compositional language
- Training Quantum Algorithms means training algorithms way closer to hardware

References

[1] Ashwani, S., Hegde, K., Mannuru, N. R., Jindal, M., Sengar, D. S., Kathala, K. C. R., ... & Chadha, A. (2024). Cause and Effect: Can Large Language Models Truly Understand Causality?. *arXiv preprint arXiv:2402.18139*.

[2] Carey, Alycia & Wu, Xintao. (2022). The Fairness Field Guide: Perspectives from Social and Formal Sciences.

[3] Khetan, V., Ramnani, R., Anand, M., Sengupta, S., & Fano, A. E. (2022). Causal bert: Language models for causality detection between events expressed in text. In *Intelligent Computing: Proceedings of the 2021 Computing Conference, Volume 1* (pp. 965-980). Springer International Publishing.

[4] Kıcıman, E., Ness, R., Sharma, A., & Tan, C. (2023). Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

[5] Long, S., Schuster, T., & Piché, A. (2023). Can large language models build causal graphs?. *arXiv preprint arXiv:2303.05279*.

References

[6] Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

[7] Tull, S., Lorenz, R., Clark, S., Khan, I., & Coecke, B. (2024). Towards Compositional Interpretability for XAI. *arXiv preprint arXiv:2406.17583*.

[8] Zhang, C., Bauer, S., Bennett, P., Gao, J., Gong, W., Hilmkil, A., ... & Vaughan, J. (2023). Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint arXiv:2304.05524*.

References (QNLP)

[1] Guarasci R, De Pietro G, Esposito M. Quantum Natural Language Processing: Challenges and Opportunities. *Applied Sciences*. 2022; 12(11):5651. https://doi.org/10.3390/app12115651 (https://www.mdpi.com/2076-3417/12/11/5651)

[2] Garg and Ramakrishnan 2020

[3] Ganguly et al. 2022

[4] Lewis 2019

[5] https://pennylane.ai/qml/glossary/circuit_ansatz

[6] https://www.nature.com/articles/s41534-019-0223-2.pdf

[7] https://docs.strangeworks.com/apps/qaoa

[8] Zhang et al. 2020

Study Approach

Minimal

- Work with the Slides

Standard

- Work with the Slides + Read into the first chapter of Tull et al. (2024) and Khetan et al. (2022)

In-Depth

- Standard Approach + more into Tull et al. (2024)