An introduction to quantum natural language processing and a study case

Javier Orduz [1 2] **Shubhangi Rastogi** [3] **Erich Baker** [2]

Abstract

This document provides essential concepts, materials, and examples in Quantum Natural Language Processing (QNLP). QNLP is a relatively recent topical area of research and the literature is limited. However, to reinforce this emerging literature, we review here main concepts fundamental to QNLP. Aligning applied quantum computing with NLP, with a focus on illustrating basic concepts and implementations, will provide a strong foundational framework for exercising future growth areas, and exposing potential QNLP applications.

1. Introduction

Classical computing relies on the Turing Machine model, which is sufficient for contemporary cumputing machinary. However, the proposal articulated by Feynman (Feynman, 1999), and formalized by D. Deutsch (Deutsch, 1985), referred to commonly as Quantum computing (QC), is expanding daily to different computing areas. In particular, we discuss here some concepts, implementations, and examples of Quantum Natural Language Processing (QNLP). In concert with other QC subdisciplines, QNLP represents one way in which QC will improve classical computing and support areas related to cybersecurity, health, finance, and communications.

For the purposes of reviewing basic concepts in QNLP, we present quantum ideas and QNLP topics starting from tradional, well-founded issues. In addition, we share relevant concepts from classical computation and ways in which they integrate with quantum theory. For example, we include a brief discussion of the quantum circuits model. Finally, we share a quantum circuit application within the context of QNLP. While, this document contains research in progress, we expect to show other examples in a forthcoming full version of this paper.

In quantum mechanics, physicists have a particular notation to represent a state $|\psi\rangle$ (*ket-psi*), referred to as Dirac or bracket notation. This vector lives in a complex space, and is amiable to matrix representation. In quantum computing, this state is called a qubit or qbit (Nielsen & Chuang, 2002).

Quantum computing takes quantum mechanics' entanglement, teleportation, and superposition concepts and enhances classical computing (Ayoade et al., 2022; Orduz et al., 2021). We find states that can not be written as two separate qubit states in quantum computing. We call these states entangled states. Teleportation is the phenomenon that uses the no-cloning theorem to say that we can not simply make an exact copy of an unknown quantum state. For example, we can not copy a state vector in superposition. And superposition means that we can get a state written on a (computational) basis. In addition, we use quantum mechanics notation within computing. Thus we have qubits instead of bits, a vector state instead of a simple state. And we formalize everything by introducing new operations such as tensor product, Hilbert space, and the classical concepts.

Natural language processing (NLP) is a multipronged area of research that various applications have used to process human language. It narrows the gap between machines and human language. While the nuance of human language in its native state is extremely difficult to decipher, social media platforms have made the task more complex because of (1) voluminous data, (2) misspelled words, (3) abbreviations, and (4) unstructured data. Several studies have attempted to alleviate these challanges by leveraging NLP to convert unstructured data on social media to a structured form. Furthermore, natural language understanding (NLU) and natural language generation (NLG) are two broad areas of NLP. NLU deals with the text's context, syntax, and semantics and converts it into machine-readable form. In contrast, NLG generates text that is understandable by a human.

NLP offers techniques such as lemmatization, stemming, part of speech tagging, and sentence breaking to check

[1] Department of Mathematics and Computer Science, Earlham College, 801 W Nat'l Rd, Richmond, IN 97141, USA [2] Department of Computer Science, Baylor University, One Bear Place, TX 76798, USA [3] Department of Computer Science & Engineering, Punjab Engineering College, Chandigarh, India . Correspondence to: Javier Orduz <orduzja@earlham.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

grammatical errors. Also, a few methods for semantic analysis, namely, named entity recognition (NER), word sense disambiguation, among others.

The rule-based methods are popular in NLP, but many subfields of NLP are now using machine learning-based approaches. The real-world applications of NLP involve sentiment analysis, intent mining, fake text identification, etc. Machine-learning-based procedures require a lot of time to learn the structure and grammar of the text. The amount of data (textual data) available on social media is enormous; thus, it requires powerful resources also. But quantum computing offers a different language model to process grammar as mathematical rules. Quantum computing improves the learning capability and is a state-of-the-art technique. Now, we will comprehend the theoretical benefits of NLP with quantum computing.

Quantum NLP is based on mathematical theories to connect the semantic information of words with the syntactic structure of a sentence. This area has a few papers such as (Miranda et al., 2021; Lorenz et al., 2021), and some documentation package (Kartsaklis, D., et. al. , 2021). It first parses a given sentence and converts it into a parse tree using a statistical Combinatory Categorial Grammar (CCG) parser. The parse tree is further converted into a string diagram that expresses the grammatical structure of the text. Section 2 explains the methods and models for QNLP.
However, the scalability of available quantum computers is limited. Thus, the combined approach of classical and quantum NLP operations has shown promising results in real-life applications. The hybrid model does most of the computations on classical computers, while the intensive computational operations are performed on quantum computers for effective utilization of resources. In a nutshell, the paper covers three dimensions in the field of natural language processing meeting quantum computing: **D1** NLP vs. QNLP: The paper introduces NLP and the potential transition into quantum-inspired NLP. **D2** Quantum inspired model for NLP: Quantum computing can be leveraged to model language features. **D3** Application areas of QNLP: Quantum inspired NLP has different application areas.

This paper is organized as follows, section 2, we leave models and methods in the QNLP context. Section 3, we share our preliminary results.And section 4, we discuss about potential QNLP applications.

2. Models and Methods

DisCoCat model means Distributional Compositional Category, and was proposed by (Coecke et al., 2010). The model combines embedded words along the sentence's grammatical structure to encode its meaning.

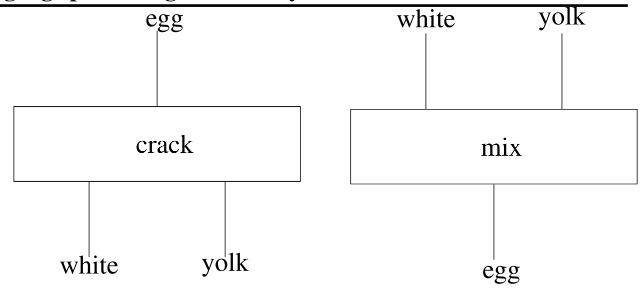

Figure 1. Split diagram with egg as input and output, and after cracking the split white and yolk as output.

To analyze the different structures in the sentences, we use Lambek rules (Lambek, 1958) and the `lambeq` package in python (Kartsaklis, 2021; Kartsaklis, D., et. al. , 2021). We start with basic examples. We will create graphs and show that we can pursue a similar methoddology in quantum circuits. While ruddimentry, the exercise works to explain fundamental concepts within a Natural Language Processing (NLP) context, and allows it to connect with the Quantum NLP. We start by taking, cracking, and separating an egg as follows (fig. 1).

The previous figure 1 shows one process with different outputs; however, if we join both charts, we obtain input as output.

Consider both diagrams1 as diagram a.k.a. string diagram, tensor networks (Biamonte, 2019) Penrose notation (Penrose, 1971; Penrose, Roger, 1972), where it shows the egg as input-output fig. 2.

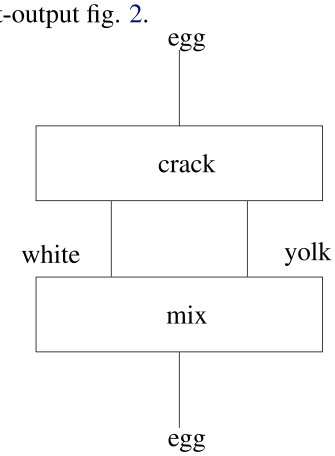

Figure 2. Diagram with the input and output egg.

In the classical context of NLP, we find the following concepts. **Monoidal category** is a directed graph with multiple edges. It is formally a collection of objects and morphisms. **String diagram** is a diagrammatic representation that reflects computations in a monoidal category. This abstraction is well-suited to model how a quantum computer works and processes data. **Tensor network** is a directed acyclic graph expressing a (multi-)linear computation between tensors (Biamonte, 2019). The graph's vertices are multi-linear tensor maps, and the edges correspond to vector spaces. **Ansatz** (plural: ansätze) this is a map between the number of qbits and the parameterized quantum states. **Combinatory Categorial Grammar (CCG)** is a grammar formalism inspired by combinatory logic and developed by Mark Steedman (Steedman, 2000). **IQP circuit** means instantaneous Quantum Polynomial. It is a circuit that interleaves layers of Hadamard quantum gates with diagonal unitaries. **DisCoPy** means DIStributional COmpositional PYthon. A Python library for working with monoidal categories. **pytket** is a Python interface for the tket compiler. **tket** this stylised $t\,|\text{ket}\rangle$. A quantum software development platform produced by Cambridge Quantum (Sivarajah et al., 2020).

We are aware of the limited concepts in this document, but we consider they are enough to implement our examples and will be augmented in future work.

Regarding quantum version, we implement packages `discopy` (de Felice et al., 2021) to do NLP, `pytket` (CQC, 2022) to connect with the `Qiskit` package (IBM, 2021), and we experiment with different circuits, and sentences in this first stage. Some preliminary results has been shown in the Collaboration and Baylor AI (2022) repository.

2.1. Quantum Circuits

We leverage existing models, such as Turing Machines and Boolean circuits, to analyze the complexity in terms of computing. The Quantum logic gate, quantum circuit, or universal quantum gate model is one of the most popular quantum computing models, and represents a sequence of quantum gates (Ayoade et al., 2022; Orduz et al., 2021). These gates are reversible transformations applied to a quantum mechanical system known as a n-qubit register (Kitaev, 1997b). The graphical depiction of quantum circuit elements is described using a variant of the Penrose graphical notation (Penrose, 1971). A quantum circuit represents a quantum operation that performs sequential calculations. Logic qubits are transported on wires (horizontal lines figs. 3 and 4), and quantum gates act on the qubits in a typical quantum circuit. The logical gate is a device that controls or processes data; the Hadamard (H) and NOT (X) gates are two common examples.

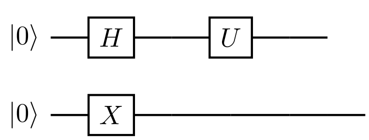

Figure 3. Quantum circuit example with Hadamard, Controlled-U gates. Qubits are initialized on 0.

A Quantum circuit is a sequence of quantum gates, measurements, and initializations of qubits that expresses a computation in a quantum computer. The purpose of `lambeq` is to convert sentences into quantum circuits that can be evaluated on quantum hardware. As we will show below, we can introduce sentences in circuits. In particular, we split sentences and use a tokenization concept.

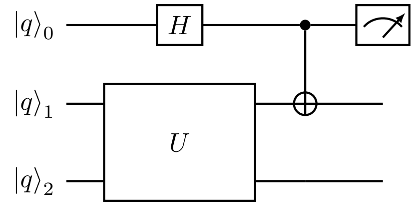

Figure 4. Quantum circuit example with Hadamard, U, Controlled-NOT, and measurement gates.

2.2. Parameterization

So far, a sentence is still represented as a string diagram, independent of any low-level decisions, such as tensor dimensions or specific quantum gate choices. We apply an ansatz to determine the number of qubits involved in each quantum state and word. An ansatz is a map, and `lambeq` can include ansatz through the classes such as `TensorAnsatz` or `CircuitAnsatz` depending on the type of the experiment.

The circuit that we are using consists of five qubits with random entanglements. Fig. 5 shows one quantum circuit obtained with the particular sentence. An arbitrary operator U can be represented by a universal gate model (Kitaev, 1997a) as:

$$U = e^{i\alpha} R_{\hat{n}}(\beta) R_{\hat{m}}(\gamma) R_{\hat{n}}(\delta) \quad (1)$$

where α is a global phase, β, γ and δ are arbitrary angles, and $\hat{m}, \hat{n}$ are non-parallel real unit vectors (Nielsen & Chuang, 2002); and $R_{\hat{n}}(\theta) \equiv e^{-i\frac{\theta}{2}\hat{n}\cdot\vec{\sigma}}$, where $\vec{\sigma}(=\sigma_i, \sigma_j, \sigma_k)$ represents the Pauli matrices; and θ an arbitrary angle. The matrices representation are

$$R_{\hat{n}}(\theta) = \cos\frac{\theta}{2} I - i \sin\frac{\theta}{2}\Big(n_x\sigma_x + n_y\sigma_y + n_z\sigma_z\Big), \quad (2)$$

eq. (2) shows β, γ and δ angles which help to parametirized the quantum circuits. And in its exponential form, $R_{\hat{n}}(\theta) \equiv e^{-i\frac{\theta}{2}\hat{n}\cdot\vec{\sigma}}$, where $\hat{n} = (n_x, n_y, n_z)$,

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \text{and } \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad (3)$$

Pauli matrices have particular properties, such as, $\sigma_x\sigma_y\sigma_x = -\sigma_y$. The following section discusses results, elements, and potential applications.

3. Results

We realize different experiments and quantum circuits are more complex if we increase the number of words in the sentences. In particular, we implemented the variable,

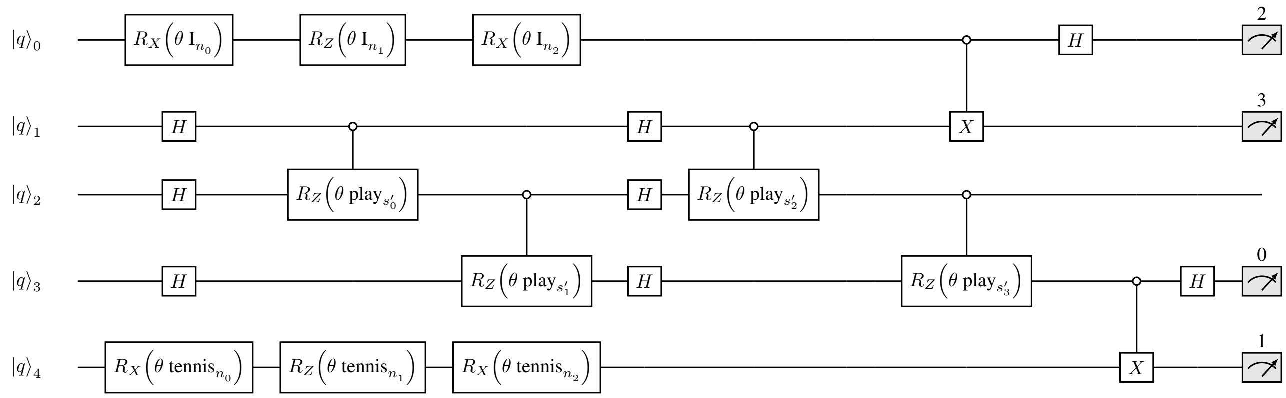

Figure 5. Encoded Circuit. Where $\theta = 2\pi$, $s' = n \cdot r@s@n \cdot l$

```
sentence = 'I play tennis'
```

and we obtained the circuit shown in fig.5. We realized different experiments, varying the number of layers and the sentence. We obtained quantum circuits that were more complex to increase the number of elements in the sentence.

We expect to incorporate some applications in the subsequent tests and implement new concepts and solutions in the next stage. In addition, we expect this field can be explored in the next years and provides new areas of knowledge.

4. Discussion

We embedded words in the phase of quantum circuits, in particular, $R_{X,Z}$ gates. These gates store information inside the function argument and then apply other gates to share the information with high security, reduce storage, or improve the computer performance. However, more algorithms and applications need to be explored since they currently have several engineering developments and implementation limitations.

Potential QNLP applications are numerous. QNLP has outperformed the strong basic models for several application areas:

Language translation: A strong literature exists in machine translation. Researchers have explored several classical NLP techniques using neural models, encoder-decoder models, encoder-decoder with attention, and word embedding for machine translation. There is still a need for error-free machine translation models. Unlike classical models, quantum computers build the network of words in a sentence to obtain the exact meaning and context. (Vicente Nieto, 2021) explored this area of quantum machine translation by expressing the language mathematically. Due to the high computation power of quantum computers, it can improve machine translation.

Exploration of fake news: Fake news detection is an emerging research area, yet challenging due to the explosion of information on online platforms. The literature highlights several solutions to automate the process of fake news detection using artificial intelligence. (Zhang et al., 2019) worked on a text classification problem to analyze the sentiments of a conversation. The existing NLP approaches disregard the dependency relationships between users involved in the conversation. Thus, they employed quantum computing to model complex mathematical formulation, which was further fed into the Long short-term memory (LSTM) model. Therefore, the quantum-inspired model can be a potential solution for application areas like fake news detection.

Pre-processing of database: Machine learning models require preprocessing techniques to convert raw data into machine-compatible and structured form. Classical NLP and machine learning-based preprocessing techniques are mean normalizations, feature scaling, dimensionality reduction techniques, etc. (Soni et al., 2020) highlighted one important preprocessing technique for dimensionality reduction, i.e., Principal component analysis (PCA). It keeps only dimensions with correlation importance and ignores the irrelevant ones. But it takes polynomial time, which can be handled by parallel quantum computers to improve the computational speed. Data preprocessing is another potential application area of quantum computing.

Speech Recognition: introduced a quantum neural network for speech recognition. Quantum neural networks have shown better arithmetic results, and fast computational speed (Fu & Dai, 2009).

This is a research in progress, and more examples and applications will be reported soon elsewhere.

References

Ayoade, O., Rivas, P., and Orduz, J. Artificial intelligence computing at the quantum level. *Data*, 7(3), 2022. ISSN 2306-5729. doi: 10.3390/data7030028. URL https://www.mdpi.com/2306-5729/7/3/28.

Biamonte, J. Lectures on quantum tensor networks, 2019. URL https://arxiv.org/abs/1912.10049.

Coecke, B., Sadrzadeh, M., and Clark, S. Mathematical foundations for a compositional distributional model of meaning. 2010. doi: 10.48550/ARXIV.1003.4394. URL https://arxiv.org/abs/1003.4394.

Collaboration and Baylor AI. QNLP, 2022. URL https://github.com/jaorduz/QNLP. Check the github repository.

CQC, 2022. See Pytket repository, https://cqcl.github.io/tket/pytket/api/.

de Felice, G., Toumi, A., and Coecke, B. DisCoPy: Monoidal categories in python. *Electronic Proceedings in Theoretical Computer Science*, 333: 183–197, feb 2021. doi: 10.4204/eptcs.333.13. URL https://doi.org/10.4204%2Feptcs.333.13.

Deutsch, D. Quantum theory, the Church-Turing principle and the universal quantum computer. *Proceedings of the Royal Society of London Series A*, 400(1818):97–117, July 1985. doi: 10.1098/rspa.1985.0070.

Feynman, R. P. *Simulating Physics with Computers*, pp. 133–153. Perseus Books, USA, 1999. ISBN 0738200573.

Fu, L. and Dai, J. A speech recognition based on quantum neural networks trained by ipso. In *2009 International Conference on Artificial Intelligence and Computational Intelligence*, volume 2, pp. 477–481. IEEE, 2009.

IBM. Qiskit: An open-source framework for quantum computing, 2021.

Kartsaklis, Dimitri, e. a. lambeq: An efficient high-level python library for quantum nlp, 2021. URL https://arxiv.org/abs/2110.04236. All authors: Kartsaklis, Dimitri and Fan, Ian and Yeung, Richie and Pearson, Anna and Lorenz, Robin and Toumi, Alexis and de Felice, Giovanni and Meichanetzidis, Konstantinos and Clark, Stephen and Coecke, Bob.

Kartsaklis, D., et. al. . lambeq: An Efficient High-Level Python Library for Quantum NLP, 2021. URL https://github.com/CQCL/lambeq. All authors: Kartsaklis, D., and Fan, I., and Yeung, R., and Pearson, A., and Lorenz, R., and Toumi, A., and deFelice, G., and Meichanetzidis, K., and Clark, S., and Coecke, B.

Kitaev, A. Y. Quantum computations: algorithms and error correction. *Russian Mathematical Surveys*, 52(6):1191–1249, dec 1997a. doi: 10.1070/rm1997v052n06abeh002155. Check https://doi.org/10.1070/rm1997v052n06abeh002155.

Kitaev, A. Y. Quantum computations: algorithms and error correction. *Uspekhi Matematicheskikh Nauk*, 52(6):53–112, 1997b.

Lambek, J. The mathematics of sentence structure. *The American Mathematical Monthly*, 65(3): 154–170, 1958. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2310058.

Lorenz, R., Pearson, A., Meichanetzidis, K., Kartsaklis, D., and Coecke, B. Qnlp in practice: Running compositional models of meaning on a quantum computer, 2021. URL https://arxiv.org/abs/2102.12846.

Miranda, E. R., Yeung, R., Pearson, A., Meichanetzidis, K., and Coecke, B. A quantum natural language processing approach to musical intelligence, 2021. URL https://arxiv.org/abs/2111.06741.

Nielsen, M. A. and Chuang, I. Quantum computation and quantum information, 2002.

Orduz, J., Rivas, P., and Baker, E. Quantum Machine Learning Foundations and Applications: A Succinct Literature Review. In *Transactions on Computational Science and Computational Intelligence*. Springer, Jul 2021. Accepted, to be published soon.

Penrose, R. Applications of negative dimensional tensors. *Combinatorial mathematics and its applications*, 1:221–244, 1971. See PDF: https://www.mscs.dal.ca/ selinger/papers/graphical-bib/public/Penrose-applications-of-negative-dimensional-tensors.pdf.

Penrose, Roger. On the nature of quantum geometry, 1972. Magic Without Magic, Freeman, San Francisco. https://ncatlab.org/nlab/files/PenroseQuantumGeometry.pdf.

Sivarajah, S., Dilkes, S., Cowtan, A., Simmons, W., Edgington, A., and Duncan, R. t|ket⟩: a retargetable compiler for NISQ devices. *Quantum Science and Technology*, 6(1):014003, nov 2020. doi: 10.1088/2058-9565/ab8e92. URL https://doi.org/10.1088/2058-9565/ab8e92.

Soni, B., Khare, N., Soni, K. K., and Rasool, A. Classical equivalent quantum based efficient data preprocessing algorithm. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–7. IEEE, 2020.

Steedman, M. *The Syntactic Process*. MIT Press, Cambridge, MA, USA, 2000. ISBN 0262194201.

Vicente Nieto, I. Towards machine translation with quantum computers, 2021.

Zhang, Y., Li, Q., Song, D., Zhang, P., and Wang, P. Quantum-inspired interactive networks for conversational sentiment analysis. pp. 5436–5442, 7 2019. doi: 10.24963/ijcai.2019/755. URL https://doi.org/10.24963/ijcai.2019/755.