Language, Information, and Quantum Theory

Thomas L. Draper

A senior thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Bachelor of Science

Jean-Francois Van Huele, Advisor

Department of Physics and Astronomy

Brigham Young University

ABSTRACT

Language, Information, and Quantum Theory

Thomas L. Draper
Department of Physics and Astronomy, BYU
Bachelor of Science

This thesis introduces concepts from quantum theory, information theory, and linguistics, and explores the connections between these fields. This culminates in a discussion of the DisCoCat model of language and the implications of quantum models of meaning on the distribution of information content in language.

Keywords: computing, entanglement, entropy, information, language, quantum theory

ACKNOWLEDGMENTS

I would first like to acknowledge my advisor Jean-François Van Huele, who has provided helpful guidance, insights, discussion, and feedback. Additionally, I acknowledge financial support from the College of Physical and Mathematical Sciences at BYU. The idea for this thesis grew out of a final project for a natural language processing class taught by Deryle Lonsdale.

Contents

Chapter 1

Introduction

In Chapter 2, we go over some fundamental concepts in quantum theory necessary for understanding this thesis. In Chapter 3, we briefly explore information theory, both classical and quantum. Then, in Chapter 4, we investigate methods of processing language, as well as connections with information and quantum theory. We use Qiskit to directly implement the DisCoCat model of sentence meaning for a small dataset. Much of the background material on quantum computation and quantum algorithms for language processing is borrowed from my recent paper [1]. Lastly, in Chapter 5, we use a simplified quantum model of language meaning to explore applications of quantum information theory to language.

Chapter 2

Quantum Theory

The universe follows familiar patterns, such as how a resting object doesn't change its motion unless a force is applied. Classical computers use these laws to store, transfer, and process information. However, we know that, at a microscopic scale, the universe also follows quantum rules which are quite different from the classical rules. Devices based on these microscopic physical laws are theorized to perform some computations asymptotically faster than classical computers, and their unique information processing capabilities may even be able to magnify the power of AI. This chapter introduces the basics of quantum theory necessary for the rest of this thesis. For additional information, Nielsen and Chuang's book [2] is an excellent reference.

2.1 Qubits and Dirac Notation

The Schrödinger equation is commonly expressed as a linear differential equation, which means that given any two solutions $\psi_1(x)$ and $\psi_2(x)$, their linear combinations $\alpha\psi_1(x) + \beta\psi_2(x)$, or superpositions, are also solutions. Since the equation is complex-valued, these α and β coefficients are also allowed to be complex-valued, and in fact the set of solutions to the equation forms a complex vector space. There are interesting wave properties of these spatial solutions $\psi(x)$, but for the purpose of this thesis, it will suffice to know that states are represented by vectors, and operations are represented by length-preserving (unitary) linear transformations.

To simplify the theory and practical concerns, quantum computation focuses on the case where a system has a 2-dimensional solution space, in which case we call the system a qubit. We can take two orthogonal basis vectors and write them suggestively as $|0\rangle$ and $|1\rangle$, for analogy with classical computing, where the fundamental unit is the bit, which can be in one of two states, 0 or 1. The "ket" symbol $|\psi_1\rangle$ represents a vector in Dirac notation, and the corresponding dual vector is represented with a "bra" $\langle\psi_2|$, so that an inner product can be represented as a "bra-ket" $\langle\psi_2|\psi_1\rangle$. In physics, we use the convention that physically realizable states for a system have norm 1 under the Hilbert space (complex vector space) norm, so we might as well take these basis vectors to have norm 1: $|\,|0\rangle\,| = |\,|1\rangle\,| = 1$. This then means that any possible state of the qubit will be $\alpha\,|0\rangle + \beta\,|1\rangle$ where $|\alpha|^2 + |\beta|^2 = 1$. We may interpret $|\alpha|^2$ or $|\beta|^2$ as respective probabilities of the system collapsing into states $|0\rangle$ or $|1\rangle$ upon measurement. This interpretation is known as the Born rule.

Not only is the description of a single qubit state broader than that of a classical bit, but qubits also combine together in a more subtle manner. For a 3-bit classical system, we would say that each bit is either in the state 0 or in the state 1, so the system is in one of $2^3 = 8$ possible states: 000, 001, 010, 011, 100, 101, 110, or 111. When combining 2-state quantum systems, their vector spaces combine in the natural way, a tensor product; given a collection of vector spaces, their tensor product is a new vector space whose basis elements are identified by choosing one basis element from each of the input vector spaces. Since each subsystem can be in its own state $|0\rangle$ or $|1\rangle$, we can similarly write $|a\rangle\,|b\rangle\,|c\rangle$ or $|abc\rangle$ to indicate that the first subsystem is in state $|a\rangle$, the second in state $|b\rangle$, and the third in state $|c\rangle$. So in this 3-qubit example we can write eight basis states $|000\rangle$, $|001\rangle$, $|010\rangle$, $|011\rangle$, $|100\rangle$, $|101\rangle$, $|110\rangle$, $|111\rangle$, and then the possible states for the entire system will be the normalized (norm 1) linear combinations of these $2^3 = 8$ basis states. Independent systems which aren't in basis states can be combined by using the linearity of the tensor product. For example, if the first system is in state $|\psi_1\rangle = \frac{1}{\sqrt{3}}\left(|0\rangle + \sqrt{2}\,|1\rangle\right)$ and the second system is in state $|\psi_2\rangle = \frac{1}{\sqrt{3}}\left(\sqrt{2}\,|0\rangle + i\,|1\rangle\right)$, then the composite system is in the state

$$\begin{aligned}
|\psi_1\psi_2\rangle &= |\psi_1\rangle\,|\psi_2\rangle \\
&= \frac{1}{\sqrt{3}}\left(|0\rangle + \sqrt{2}\,|1\rangle\right)\frac{1}{\sqrt{3}}\left(\sqrt{2}\,|0\rangle + i\,|1\rangle\right) \\
&= \frac{1}{3}\left(\sqrt{2}\,|0\rangle\,|0\rangle + 2\,|1\rangle\,|0\rangle + i\,|0\rangle\,|1\rangle + \sqrt{2}i\,|1\rangle\,|1\rangle\right) \\
&= \frac{1}{3}\left(\sqrt{2}\,|00\rangle + 2\,|10\rangle + i\,|01\rangle + \sqrt{2}i\,|11\rangle\right).
\end{aligned}$$

Measurements probabilities in the standard basis can be performed by taking the squared norm of the corresponding coefficient. For instance, in the previous example $|\psi_1\psi_2\rangle$, the probability of measuring $|11\rangle$ would be $|\sqrt{2}i/3|^2 = 2/9$. Measurements in other bases are most easily represented by applying bra dual vectors to the given ket vector, expanding by linearity, and applying the orthogonality relations $1 = \langle 00|00\rangle = \langle 01|01\rangle = \cdots$ and $0 = \langle 00|01\rangle = \langle 00|10\rangle = \cdots$. For example, if we measured the state above in the basis of $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ and $|-\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$, then the probability of measuring $|{+}{+}\rangle$ would be

$$\begin{aligned}
|\langle{+}{+}|\psi_1\psi_2\rangle|^2 &= \left|\frac{1}{2}\Big(\langle 00| + \langle 01| + \langle 10| + \langle 11|\Big)\frac{1}{3}\left(\sqrt{2}\,|00\rangle + 2\,|10\rangle + i\,|01\rangle + \sqrt{2}i\,|11\rangle\right)\right|^2 \\
&= \left|\frac{1}{6}\left(\sqrt{2} + 2 + i + \sqrt{2}i\right)\right|^2 \\
&= \frac{3 + 2\sqrt{2}}{12}.
\end{aligned}$$

2.2 Bell Tests of Entanglement and the CHSH Inequality

The tensor product structure for composition of qubit systems allows for interesting kinds of correlations between different qubits. In this section, we restrict our attention to two-qubit systems, which provide the simplest example of quantum entanglement. The famous Bell state can be written in Dirac notation $\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. In this section, we explore what it means that this state is maximally entangled.

Upon measuring the Bell state in the standard basis, there is a 50% chance that both qubits come up as 0, and 50% chance that both come up as 1. It is true that this one distribution of measurements doesn't require any quantum mechanics, and in fact, if we just prepared states where 50% of them are $|00\rangle$ and 50% are $|11\rangle$, we would get identical measurement outcomes in the standard basis. The significance of entanglement lies in how the system evolves, and in particular how measurements remain correlated, even when the system is measured in different bases. For example, if we switch to the basis consisting of $|+\rangle$ and $|-\rangle$, then the Bell state is just $\frac{1}{\sqrt{2}}(|{++}\rangle + |{--}\rangle)$, so our measurement outcomes would be 50% $|{++}\rangle$ and 50% $|{--}\rangle$. We contrast this with the classical probability distribution of 50% $|00\rangle = \frac{1}{2}(|{++}\rangle + |{+-}\rangle + |{-+}\rangle + |{--}\rangle)$ and 50% $|11\rangle = \frac{1}{2}(|{++}\rangle - |{+-}\rangle - |{-+}\rangle + |{--}\rangle)$, in which case we would measure 25% of each of $|{++}\rangle$, $|{+-}\rangle$, $|{-+}\rangle$, and $|{--}\rangle$, a uniform mixture.

The difference between classical probability distributions and distributions arising from measurements of coherent entangled quantum states can be mathematically quantified, as Bell proved [3]. Bell's theorem can be expressed using the CHSH inequality [4], which is commonly presented as follows. Let $A = \{a_0, a_1\}$ and $B = \{b_0, b_1\}$ be two single-qubit orthogonal measurement bases, and let $p(a_i, b_j)$ represent the probability of measuring both a_i on the first qubit and b_j on the second qubit when the two qubits are measured in basis A and B, respectively. Then we can define a measure of correlation:

$$E(A,B) := p(a_0,b_0) - p(a_0,b_1) - p(a_1,b_0) + p(a_1,b_1).$$

We note that $-1 \leq E(A,B) \leq 1$ since the four probabilities are nonnegative and add to one. The Bell state $\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$ will give $\frac{1}{2} - 0 - 0 + \frac{1}{2} = 1$ in the standard basis, which indicates the maximum possible correlation, but recall that quantum entanglement is characterized by correlation in several measurement bases, not just one. The statement of the CHSH inequality is

$$-2 \leq E(A,B) - E(A,B') + E(A',B) + E(A',B') \leq 2,$$

where A, A', B, and B' are arbitrary measurement bases. The CHSH inequality is based on the assumption of local realism, meaning that two separated systems cannot affect each other instantaneously, and that physical systems are always in a definite state, even if we don't know what that state is or are not observing it. Absent any assumptions, we could get arbitrary values between -4 and 4, but under the assumptions of quantum mechanics using orthogonal measurements, we instead get Tsirelson's bound [5],

$$-2\sqrt{2} \leq E(A,B) - E(A,B') + E(A',B) + E(A',B') \leq 2\sqrt{2}.$$

This bound can be achieved on the Bell state by cleverly choosing the measurement bases so that three of the correlation measures are $\frac{1}{\sqrt{2}}$, while the last is $-\frac{1}{\sqrt{2}}$. Many experiments have been performed on photon pairs which are separated and then measured in different polarization bases, and the results agree with Tsirelson's bound and violate the CHSH inequality. This shows that quantum mechanics accurately describes correlations between physical systems in a way that necessarily violates local realism.

Lastly, we consider the concept of entanglement in more detail. A state on two quantum systems is said to be entangled if it cannot be written as the product of two states, one for each subsystem. For example, $\frac{1}{\sqrt{2}}(|00\rangle + |01\rangle) = |0\rangle \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ is not entangled, but the Bell state is; no matter how hard you may try, you cannot write it as a product. On a product state, the measurement of each subsystem is independent, meaning that $p(a_i, b_j) = p(a_i)p(b_j)$, using the notation as in the definition of $E(A,B)$, where $p(a_i)$ represents the (marginal) probability of measuring a_i on system A. For qubits, we only have two states, so $p(a_1) = 1 - p(a_0)$ and $p(b_1) = 1 - p(b_0)$, so on a product state,

$$\begin{aligned} E(A,B) &= p(a_0)p(b_0) - p(a_0)(1-p(b_0)) - (1-p(a_0))p(b_0) + (1-p(a_0))(1-p(b_0)) \\ &= (1-2p(a_0))(1-2p(b_0)). \end{aligned}$$

Since the correlation E decomposes as a product, the measure in the CHSH inequality is

$$\begin{aligned} &E(A,B) - E(A,B') + E(A',B) + E(A',B') \\ &= (1-2p(a_0))[(1-2p(b_0)) - (1-2p(b_0'))] + (1-2p(a_0'))[(1-2p(b_0)) + (1-2p(b_0'))] \\ &= (1-2p(a_0))[2p(b_0') - 2p(b_0)] + (1-2p(a_0'))[2 - 2p(b_0) - 2p(b_0')]. \end{aligned}$$

Since probabilities are between zero and one, $|(1-2p(a_0))| \leq 1$, and

$$|2p(b_0') - 2p(b_0)| + |2 - 2p(b_0) - 2p(b_0')| \leq 2,$$

so we see that the CHSH inequality must always hold for a product state. Therefore, violation of the CHSH inequality requires an entangled state (and in fact, any entangled pure state violates a Bell inequality [6]), so the maximum of $|E(A,B) - E(A,B') + E(A',B) + E(A',B')|$ over possible bases can be seen as another measurement of entanglement.

2.3 Density Matrices

So far, we have only considered pure quantum states, for which we know everything about the system and write it using a ket $|\psi\rangle$. In general, there may be a probability distribution over states which our system may be in. In Section 2.2, we reasoned about correlation differences between quantum superpositions and classical probability mixtures, especially with regards to entanglement. There is actually a mathematical way to represent such mixtures, within the framework of the density matrix, or density operator. Given a mixture of $|\psi_i\rangle$s, each with probability p_i, the state is called a mixed state, and the corresponding density operator is

$$\rho := \sum_i p_i \left|\psi_i\right\rangle \left\langle\psi_i\right| .$$

For example, for a classical mixture of 50% $|00\rangle$ and 50% $|11\rangle$, we get the matrix

$$\begin{aligned}\rho &= \frac{1}{2}\left|00\right\rangle\left\langle 00\right| + \frac{1}{2}\left|11\right\rangle\left\langle 11\right| \\ &= \begin{array}{c c} & \begin{array}{cccc} \langle 00| & \langle 01| & \langle 10| & \langle 11| \end{array} \\ \begin{array}{c} |00\rangle \\ |01\rangle \\ |10\rangle \\ |11\rangle \end{array} & \left[\begin{array}{cccc} 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 \end{array}\right] \end{array} .\end{aligned}$$

For the coherent Bell state, we instead get

$$\begin{aligned}\rho &= \frac{1}{\sqrt{2}}(\left|00\right\rangle + \left|11\right\rangle)\frac{1}{\sqrt{2}}(\left\langle 00\right| + \left\langle 11\right|) \\ &= \frac{1}{2}(\left|00\right\rangle\left\langle 00\right| + \left|00\right\rangle\left\langle 11\right| + \left|11\right\rangle\left\langle 00\right| + \left|11\right\rangle\left\langle 11\right|) \\ &= \begin{array}{c c} & \begin{array}{cccc} \langle 00| & \langle 01| & \langle 10| & \langle 11| \end{array} \\ \begin{array}{c} |00\rangle \\ |01\rangle \\ |10\rangle \\ |11\rangle \end{array} & \left[\begin{array}{cccc} 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 \end{array}\right] \end{array} .\end{aligned}$$

The off-diagonal terms are called coherence terms, and this example illustrates the reason well. If you have a classical mixture of basis states, then there are no off-diagonal terms, but if you have a coherent quantum state which is in a superposition, then you get off-diagonal terms in the density matrix, corresponding to cross terms in the outer product.

One of the main mathematical benefits of using density matrices is the ability to represent states of subsystems in a consistent way. As seen in Section 2.2, there is no way to represent the Bell state as a product of two individual qubit states, and so we can't get a single vector for each subsystem. However, we can do this for density matrices, using the partial trace. The definition of the trace of a matrix is the sum of the diagonal elements:

$$\operatorname{tr}(A) := \sum_i A_{ii}.$$

For example,

$$\operatorname{tr}\left(\begin{array}{c|cccc} & \langle 00| & \langle 01| & \langle 10| & \langle 11| \\ |00\rangle & 1/2 & 0 & 0 & 0 \\ |01\rangle & 0 & 0 & 0 & 0 \\ |10\rangle & 0 & 0 & 0 & 0 \\ |11\rangle & 0 & 0 & 0 & 1/2 \end{array} \right) = 1/2 + 0 + 0 + 1/2 = 1,$$

and in fact $\operatorname{tr}(\rho) = \sum_i p_i \langle\psi|\psi\rangle = \sum_i p_i = 1$ for any density matrix ρ. A partial trace is when we add the terms corresponding to just one subsystem, leaving us with a smaller density matrix. If we write A for one qubit and B for the other, then we may call the composite system AB, and tracing out system B can be written as

$$\rho_A = \operatorname{tr}_B(\rho_{AB}) = \langle 0_B| \rho_{AB} |0_B\rangle + \langle 1_B| \rho_{AB} |1_B\rangle .$$

For example,

$$\operatorname{tr}_B \left(\frac{1}{2} |0_A 0_B\rangle \langle 0_A 0_B| + \frac{1}{2} |1_A 1_B\rangle \langle 1_A 1_B| \right) = \frac{1}{2} |0_A\rangle \langle 0_A| + \frac{1}{2} |1_A\rangle \langle 1_A| ,$$

so if we have two qubits in a classical mixture of $|00\rangle$ and $|11\rangle$, then each individual qubit can be considered a classical mixture of $|0\rangle$ and $|1\rangle$. The interesting part is that we get this same result upon applying the partial trace to the Bell state $\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. Because the system is entangled, you cannot say that the subsystems are in a particular pure state, but if you really want to ignore the other qubit, you can just partial trace it out and end up with a density matrix representing a mixed state. This partial trace operation will be useful when we consider how information is distributed within a quantum system.

2.4 Quantum Computation

Any physical operation (often called a gate in quantum computing) must map physical state inputs to physical state outputs, so in particular, it must take unit (normalized) vectors to other unit vectors. In the language of linear algebra, this is a unitary transformation, which also maps mutually orthogonal vectors to mutually orthogonal vectors. In the finite dimensional case which we deal exclusively with in quantum computing, a unitary transformation can be expressed as a unitary matrix. Since matrices naturally represent what output vector each of a given basis of input vectors gets mapped to, it is common to characterize quantum gates by where they map each standard basis state. An arbitrary such operation could be very complicated, causing difficulty both for thinking conceptually about it and for physically implementing it, so we generally focus on basic operations and composing them together. These basic gates are described just in terms of their actions on a few relevant qubits.

As a first example of a quantum gate, one might want to perform an operation on a 2-qubit system for which one of the output bits will always be the AND of the two inputs for all four possible input combinations. However, this is impossible, since $|00\rangle$, $|01\rangle$ and $|10\rangle$ would all have to map into a vector space of dimension at most 2, which is impossible for invertible (and in particular unitary) transformations. So generalizations of natural operations from classical computing are not necessarily natural operations to perform in the context of quantum computing. One acceptable quantum operation is the NOT operator which swaps

$$|0\rangle \mapsto |1\rangle \text{ and } |1\rangle \mapsto |0\rangle,$$

which also could be written in matrix form as $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

Another simple 1-qubit gate is the PHASE(θ) operator, which maps

$$|0\rangle \mapsto |0\rangle \text{ and } |1\rangle \mapsto e^{i\theta} |1\rangle$$

with corresponding matrix $\begin{bmatrix} 1 & 0 \\ 0 & e^{i\theta} \end{bmatrix}$. Note that PHASE($\theta$), unlike NOT, has no analog in digital classical computers, where there is no concept of phase.

Slightly more complicated are the “controlled” versions of these gates, which use two qubits: a control qubit and a target qubit. An operation is applied to the target qubit if the control qubit is $|1\rangle$, while no operation is applied if the control qubit is $|0\rangle$. For example, the CNOT operation (with the first qubit as the control and the second as the target) maps

$$\begin{aligned} |00\rangle &\mapsto |00\rangle, \\ |01\rangle &\mapsto |01\rangle, \\ |10\rangle &\mapsto |11\rangle, \\ \text{and } |11\rangle &\mapsto |10\rangle. \end{aligned}$$

The controlled phase gate, CPHASE(θ), is defined similarly, with the last two rows instead being $|10\rangle \mapsto |10\rangle$ and $|11\rangle \mapsto e^{i\theta} |11\rangle$. Recall that these gates are still just linear maps defined on our chosen basis vectors, and can therefore also be represented in matrix form. For example, the CPHASE(θ) is given by

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & e^{i\theta} \end{bmatrix}.$$

Another important gate is the Hadamard gate, taking

$$|0\rangle \mapsto \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle) \text{ and } |1\rangle \mapsto \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle),$$

or in matrix form, $H = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$, which is useful since we generally assume that the initial state of any computation is the basis state where every qubit is $|0\rangle$, and so applying the Hadamard gate to every qubit can generate an equal superposition of all possible basis states, for example,

$$\begin{aligned} H\left|0\right\rangle &= \frac{1}{\sqrt{2}}(\left|0\right\rangle + \left|1\right\rangle), \\ H^{\otimes 2}\left|00\right\rangle &= \frac{1}{2}(\left|00\right\rangle + \left|01\right\rangle + \left|10\right\rangle + \left|11\right\rangle), \\ H^{\otimes 3}\left|000\right\rangle &= \frac{1}{2\sqrt{2}}(\left|000\right\rangle + \left|001\right\rangle + \left|010\right\rangle + \left|011\right\rangle + \left|100\right\rangle + \left|101\right\rangle + \left|110\right\rangle + \left|111\right\rangle). \end{aligned}$$

Lastly, we introduce x and z rotation gates parameterized by an "angle" θ, which, by analogy with 3-dimensional space, should together be able to rotate a unit vector to any other unit vector:

$$R_x(\theta) = \begin{bmatrix} \cos(\theta/2) & -i\sin(\theta/2) \\ -i\sin(\theta/2) & \cos(\theta/2) \end{bmatrix} \text{ and } R_z(\theta) = \begin{bmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{bmatrix}.$$

There are also controlled versions of these gates, for example,

$$CR_z(\theta) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & e^{-i\theta/2} & 0 \\ 0 & 0 & 0 & e^{i\theta/2} \end{bmatrix}.$$

These are the gates that we apply in Chapter 4 to represent meaning in language. We will see that parameterized gates can represent word meanings learned with a machine learning algorithm, and the CNOT and Hadamard gates can represent grammatical relationships.

Chapter 3

Information Theory

The physical resource requirements for representing information can be quantified using entropy. The concept of entropy originated in classical thermodynamics, where energy can be used to heat objects, while the original energy cannot be fully recovered as work from the heated system. This can be interpreted as a kind of disorder, where the energy of a heated object is randomized to the point that it cannot be extracted in an orderly form, such as mechanical work. Another perspective is that entropy counts the possible states that a system might be in. Given some macroscopic specification of a system, there are many possible microstates it could be in, like how our macroscopic observation of air (volume, pressure, or temperature) doesn't depend on the exact positions of each individual molecule, as long as they are roughly evenly distributed. This idea of counting possibilities ties very naturally to information theory, where we are interested in any kind of information, not just information about physical systems.

3.1 Classical Information

Claude Shannon invented information theory when he worked out a framework to determine the most efficient way to send a message [7]. The Shannon entropy is defined as

$$H(X) := H(p_1, \ldots, p_n) := -\sum_x p_x \log p_x$$

where X is a discrete random variable taking n distinct values each with probability p_i. Shannon proved that this is, up to a multiplicative constant, the unique function which satisfies three properties. First, H is a continuous function of the p_is. Second, $H(1/n, \cdots, 1/n)$ is a monotonically increasing function of positive integers n. Lastly, H can be broken down into a weighted sum corresponding to a decision tree generating the probability distribution p_i. For example, $H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}) = H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2}H(\frac{2}{3}, \frac{1}{3})$, since the distribution can be obtained by first flipping a coin to see if the first option will be taken, and if not, then there is a 2 : 1 split between the remaining two possibilities.

The standard unit of entropy is the bit, which corresponds to taking the logarithm with base 2. For example,

$$H\left(\frac{1}{2}, \frac{1}{2}\right) = -\sum_{i=0}^{1} \frac{1}{2} \log \frac{1}{2} = \log 2 = 1 \text{ bit.}$$

This makes sense, because the information necessary to determine a fair coin flip is just a single binary random variable. Further, if we want to store this information, then we need exactly one bit, a 0 or 1. We see that the Shannon entropy represents the classical physical resources required to represent information.

The Shannon entropy can be defined for a joint probability distribution in a natural way. The joint entropy is

$$H(X,Y) := -\sum_{x,y} p(x,y) \log p(x,y),$$

the entropy of the distribution of X and Y together, i.e. of the random variable (X,Y). The conditional entropy of X given Y is

$$H(X|Y) := H(X,Y) - H(Y) = -\sum_{x,y} p(x,y) \log p(x|y),$$

which describes the additional information from learning X if we already know Y. The mutual information content of X and Y is

$$H(X:Y) := H(X) + H(Y) - H(X,Y),$$

representing how much information the two variables share on average.

We note here some properties of the Shannon entropy. It is nonnegative because each $-p\log(p) \geq 0$. For multiple variables, we have $H(X:Y) \leq H(X)$, which corresponds to the fact that the mutual information of X and Y is just a subset of X's information. Also, $H(X) \leq H(X,Y) \leq H(X) + H(Y)$, which shows that given knowledge of X and Y, your knowledge is at least as much as if you knew only X, and the amount by which it increases when learning Y as well is bounded by the information content of Y itself. These facts are quite intuitive, but we will see that the corresponding quantum analogue does not follow the same rules.

3.2 Quantum Information

Given a quantum state described by density matrix ρ with eigenvalues λ_x, the von Neumann entropy is defined as

$$S(\rho) := -\operatorname{tr}(\rho \log \rho) = -\sum_x \lambda_x \log \lambda_x$$

which is similar to the Shannon entropy, but using the eigenvalues of the density operator for probabilities. Note that the “quantum” nature of the von Neumann entropy arises from how we mathematically manipulate density matrices, and in particular how density matrices of composite systems relate to reduced density matrices for component subsystems. Interesting differences in quantum information arise from the way in which systems compose, or fail to decompose.

We note that the von Neumann entropy is always nonnegative, and it is only zero for pure states. Also, if a composite system ρ_{AB} is a pure state, i.e. $S(\rho_{AB}) = 0$, then $S(\rho_A) = S(\rho_B)$ by Schmidt decomposition [2]. This has no analog in classical entropy, where the joint entropy can only be zero if both subsystems have zero entropy. Instead, we have subadditivity,

$$|S(\rho_A) - S(\rho_B)| \leq S(\rho_{AB}) \leq S(\rho_A) + S(\rho_B),$$

which shows that the entropy of the combined system may be more, equal to, or even less than the entropy of the two subsystems. This property is due to how the von Neumann entropy is defined, using the eigenvalues of the density matrix. The two subsystems may be impure independently, in which case their entropies add, or they may be in mixed states just because they are entangled with each other, in which case their entropies subtract. This measure of entropy on quantum systems does not seem to accord well with our classical ideas about information, so it may be best to think of it instead as a measure of disorder, impurity, or even ambiguity. In Chapter 5, we explore how von Neumann entropy may generate insights into the structure of ambiguities in language.

Chapter 4

Language

Language is a natural way for humans to reason and communicate, but it proves very difficult to formally represent, as is evident in the extensive work dedicated to natural language processing [8]. Written language is represented in terms of words and sentences decomposed as strings of characters. Spoken language instead uses phonemes or raw sound (fluctuations of air pressure over time) to represent words and sentences. Written formats of language are easiest to represent and work with, and even someone with no understanding can copy written characters. However, determining the meaning of a written text is a true challenge. Intelligent humans can reason about written as well as spoken language, but all attempts at formalizing this understanding into a computer program leave much to be desired. There are reasons to believe that the laws of quantum physics may hold the key to understanding the meaning of natural language [9]. In this chapter, we explore the quantum-native DisCoCat model of sentence meaning, following the presentation in a recent paper by a research group at Quantinuum and Oxford [10].

4.1 Computing with Language

Natural language processing (NLP) is all about using computers to process natural human languages, such as English. This is part of the intersection of linguistics and computer science, and also a major topic in artificial intelligence (AI). There are some basic tasks, like grammatical parsing, which are relatively easy, but the NLP field also contains what are considered to be some of the hardest problems in AI. Tasks that require understanding the full meaning of a complex text (such as translation from one language to another) are considered "AI-complete"; this means that if you can make an algorithm to (for example) reliably translate between languages, then you probably can solve just about any AI-related problem, such as object recognition in video or even generalizing knowledge to tackle problems never before encountered by computers. In this section, we explore sentence structure and word meaning, using parse trees and word vector embeddings, respectively.

4.1.1 Grammatical Parsing

The first algorithm to parse sentences was created by Lambek in 1958 [11]. Here, we introduce Lambek's pregroup grammar, using the conventions of the paper [10] for consistency.

Taking n to represent a noun, s to represent a sentence, and ^{L} and ^{R} to represent left and right inverses, respectively, we use the following definitions, explained in more detail later.

$$\text{Noun} \to n$$

$$\text{Intransitive verb} \to n^R(s)$$

$$\text{Transitive verb} \to n^R(s)n^L$$

$$\text{Relative pronoun ``who''} \;\to n^R(n)s^L n$$

These strings don't make sense on their own, but their combination rules match the grammatical structure of English. The rules we need are simple: the cancellations of symbols and their left or right inverses, when adjacent, namely $nn^R \to 1$, $n^L n \to 1$, and $s^L s \to 1$. The symbol 1 represents the identity, or empty string, so that, in essence, these rules tell us how to delete symbols to simplify type strings. The pregroup grammar is associative, so we can rearrange parentheses however we want, and this allows us to group symbols and their inverses together for deletion.

As the simplest example, we can consider the sentence "Romeo dies", which is a noun followed by an intransitive verb, and therefore has type $(n)(n^R s)$, which can be reduced to s by applying the rule $nn^R \to 1$. This indicates that "Romeo dies" is a grammatical sentence. Similarly, transitive verbs require both a subject and an object, or a noun on both sides, in order to form a sentence, which is why they are assigned the string $n^R s n^L$, and the sentence "Romeo loves Juliet" reduces from $(n)(n^R s n^L)(n)$ to s as a complete sentence. The most complex part of speech we will deal with is the relative pronoun "who", which takes a noun type on the left and the predicate type $n^R s$ on the right to yield a noun type result, hence the string $n^R(n)s^L n$.

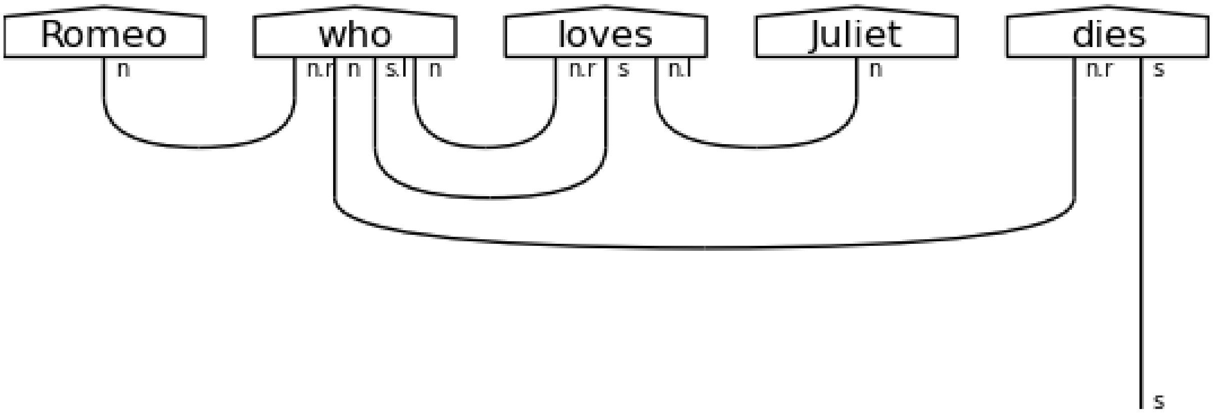

Figure 4.1 String diagram for the sentence "Romeo who loves Juliet dies", showing the grammatical connectivity of the sentence, as parsed using the pregroup grammar. This figure was generated using the `lambeq` package [12]. Note that `lambeq` draws `n.l`, `n.r`, and `s.l` to represent n^L, n^R, and s^L.

Now as our main example, we parse the sentence "Romeo who loves Juliet dies", as visualized in Figure 4.1. Here is the series of reductions:

$$
\begin{aligned}
& \texttt{(Romeo)(who)(loves)(Juliet)(dies)} && \\
\rightarrow & (n)(n^R n s^L n)(n^R s n^L)(n)(n^R s) && \text{Convert words by part of speech} \\
\rightarrow & (n n^R) n s^L (n n^R) s (n^L n) n^R s && \text{Rearrange parentheses} \\
\rightarrow & n (s^L s) n^R s && \text{Delete pairs, add parentheses} \\
\rightarrow & (n n^R) s && \text{Delete pairs, add parentheses} \\
\rightarrow & s. && \text{Delete pair}
\end{aligned}
$$

More examples of this process are provided in the papers [10, 13], and these kinds of diagrams and their applications to quantum mechanics are addressed in much greater detail in Coecke and Kissinger's book [14].

4.1.2 Word Vectors

Vectors are a computationally convenient way to compare the semantic content of words. A prototypical example of the distributional model of meaning is `word2vec` [15], where words are embedded into a vector space with the embeddings learned such that words that appear in similar contexts are close to each other (meaning that the vector space inner product between them is near 1), and words that are distributed differently in texts get embedded as vectors with small inner products. This seems to accurately describe some aspects of meaning. For example, using word vectors, if we take "king", subtract "man", and add "woman", we get something quite close to "queen".

Unfortunately, there are no good classical methods for representing entire sentence meanings. Most algorithms either work only with the grammatical structure, or ignore it entirely, as in the "bag of words" model, in which the vectors corresponding to each word in the sentence are taken without any regard to even their order. Neural networks (algorithms imitating the structure of a brain) are being used to solve NLP tasks with rapidly increasing accuracy, but these models still seem to lack an understanding of textual meaning and have difficulty generalizing to completely new contexts. If we assume that vectors are the right way to represent words, then the natural operation to perform is a multilinear map, which, as we will see, naturally lends itself to the formalism of quantum theory.

4.2 DisCoCat

DisCoCat is the name of a model of language introduced by Coecke, Sadrzadeh, and Clark [13]. It stands for **Dis**tributional **Co**mpositional **Cat**egorical, indicating the main principles behind it. Distributionality is evident in the use of vectors for words, where these vectors are learned based on actual distributions of words in text. The principle of compositionality indicates that these vectors compose together based on recursive structures, giving a formal method of building up meaning vectors for phrases and sentences. Lastly, the model is based on category theory, a mathematical framework for describing relationships and composition of mathematical structures.

4.2.1 Category Theory

A category consists of objects and associative maps between them. One example is sets and functions, but it is also common to consider categories with more structure, such as vector spaces with linear transformations. Here, we can take the tensor product between two vector spaces to construct a new vector space. This tensor operation is associative and has the identity 1-dimensional vector space $\mathbb{C}$, so it is said to be a monoid, and the category of vector spaces is a monoidal category.

In mathematics, we often often choose our ways of writing symbols based on the rules that they follow. For example, addition is associative

$$(a+b)+c = a+(b+c),$$

which justifies us in simply writing $a+b+c$ without parentheses for the sum of a, b, and c. In a similar way, the axioms of a monoidal category justify us in drawing string diagrams to represent mappings. The sentence string diagrams we used in Subsection 4.1.1 can be interpreted as maps in a monoidal category, and in particular, if we represent each word by a vector, then the string diagrams correspond to linear maps, which can also be represented by quantum circuits, which are just another kind of string diagram.

A multilinear map out of a collection of word vector spaces is equivalent to a single linear map out of the tensor product of the individual word vector spaces. This is called the universal property of the tensor product. Tensor products of word vectors with linear maps for grammatical relations allow us to compute entire sentence meanings. The dimension of a tensor product space grows exponentially in the number of spaces being combined, which means that it is not practical to try to compute sentence meanings using tensor products on classical computers.

As in the classical case, the idea of representing words as vectors is extremely natural in quantum mechanics, since a quantum system's state is always represented by a vector, so any representation of a word on a quantum computer must necessarily be a vector in some sense. Unlike in the classical case, in quantum theory, the idea that the unprocessed sentence is the tensor product of its constituent words is again so natural that it is just a byproduct of the way quantum systems are composed together. In the rest of this section, we will see how to deal with this tensor product state based on the sentence's grammatical structure.

4.2.2 Circuit Generation

In this subsection, we describe the method of transforming grammatically parsed sentences into quantum circuits, following the paper [10]. As is traditional in quantum computing, we begin with the state of all zeroes $|00\cdots00\rangle$. We can then split our circuit into three main parts: first, prepare the state corresponding to each word, next, perform the gates corresponding to the reduction of the sentence in the pregroup grammar, and finally, measure all qubits.

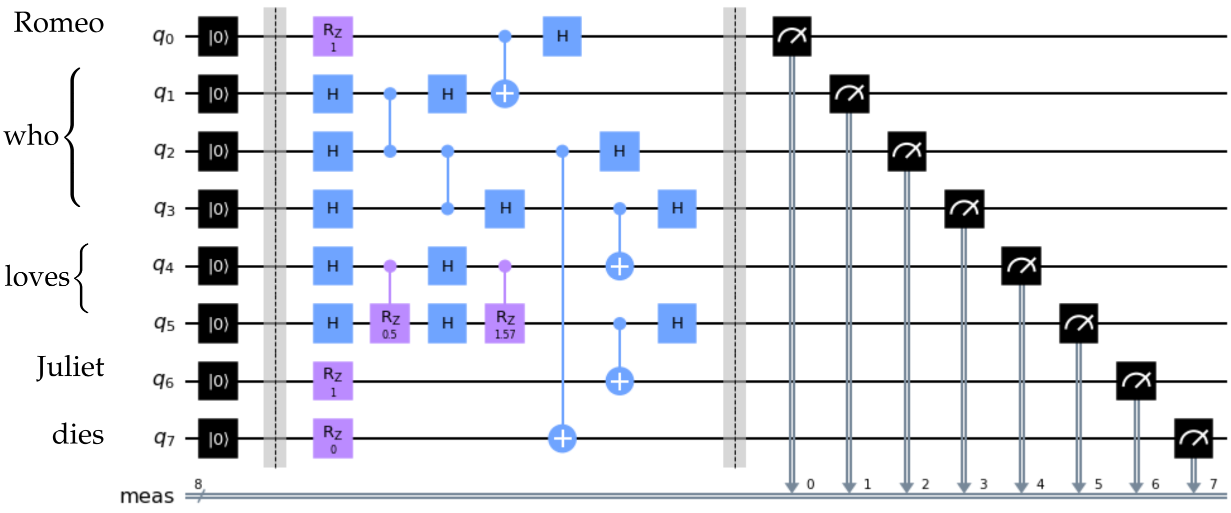

Figure 4.2 Generated quantum circuit for "Romeo who loves Juliet dies"

The circuits drawn in Figures 4.2, 4.3, and 4.4 were created in a Jupyter notebook using Qiskit [16] to run the simulations with IBM's quantum computers. These are read as follows. On the left are labels indicating which parts of the circuit correspond to which words. The whole circuit is read from left to right, showing state initialization, computational operations, and measurements. The initial q_i label the individual qubits,

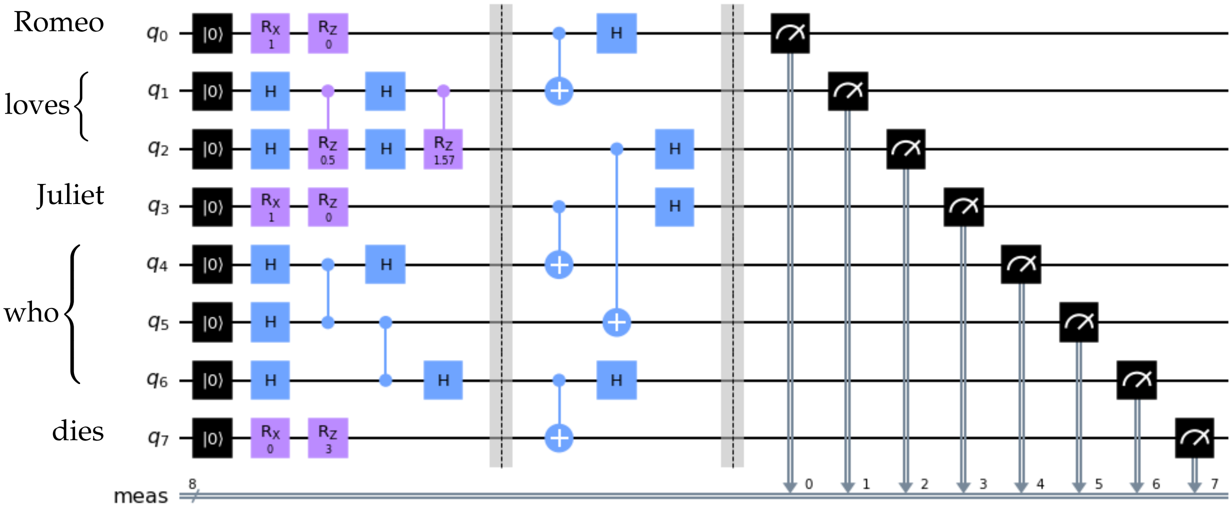

Figure 4.3 Generated quantum circuit for “Romeo loves Juliet who dies”

and the black boxes on the left show that the circuit begins in the state of all zeroes. The vertical gray lines are just cosmetic, and can be used to divide the circuit into logical groupings. Single boxes represent single-qubit gates, and colored vertical lines between wires represent multi-qubit gates. The blue `H` represents the Hadamard gate, the plain blue line represents the CPHASE(π) gate, and the blue line ending in a plus sign represents the CNOT gate, where the plus is on the target qubit. The purple `Rx` and `Rz` represent x and z rotations, respectively, and the `Rz` with a purple wire connecting to another qubit represents a controlled z rotation. Notice that these rotations are parameterized by a number, as indicated on the circuit with sample initial values. Lastly, the gauge boxes on the right side indicate a measurement, collapsing the quantum state to get some classical information about it.

The word state preparation (shown on the left part of the circuit) depends on the word’s part of speech. For nouns and intransitive verbs (like “Juliet” and “dies”), which really only have one way of relating, we quite simply put in arbitrary gates that allow any state to be made from the starting zeroes. Namely, we put in an x-rotation by θ_1 and z-rotation

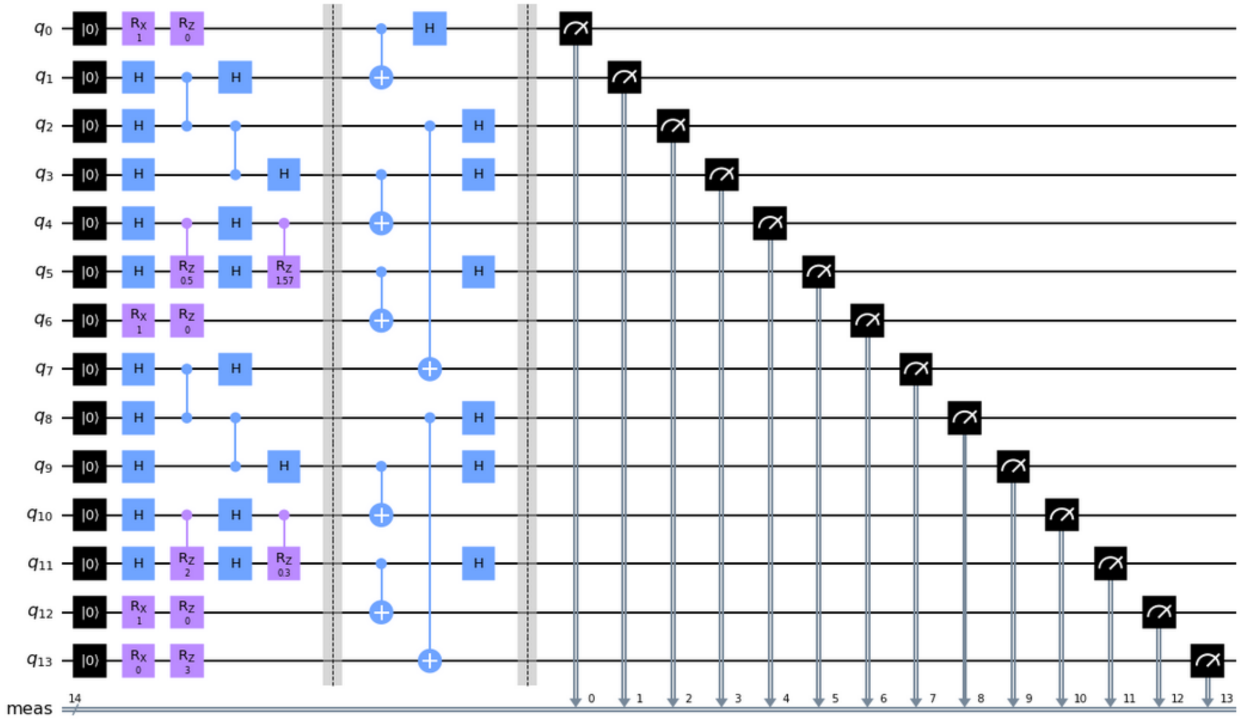

Figure 4.4 Generated quantum circuit for "Romeo (q_0) who ($q_1 - q_3$) loves ($q_4 - q_5$) Juliet (q_6) who ($q_7 - q_9$) kills ($q_{10} - q_{11}$) Juliet (q_{12}) dies (q_{13})"

by θ_2, that is $R_x(\theta_1)$ and $R_z(\theta_2)$. These parameters θ_i are yet unknown, since we don't know the exact best word embeddings, but by varying them it is possible to prepare any valid single-qubit state. For a circuit with fewer gates but a smaller embedding space, we may use a single rotation $R_z(\theta)$, as in Figure 4.2, as opposed to Figure 4.3.

For other parts of speech, the construction is a bit more complex using more qubits, but the same idea can be applied, where we have a fixed basic circuit structure for preparing the state, just with some parameters on the gates which can be varied to allow preparing arbitrary states. The transitive verb "loves" uses Hadamard gates and controlled parameterized rotations to set up an arbitrary two-qubit state representing the relationship between subject and object. Lastly, the relative pronoun "who" is assigned a series of Hadamard and CPHASE(π) gates which set up the GHZ state $\frac{1}{\sqrt{2}}(|000\rangle + |111\rangle)$, a maximally entangled three-qubit state, representing how this word ties together three different noun descriptions of the same entity. More details on the exact construction used is available in the paper [10].

Next, we need to perform some operation on this product ensuring that the different grammatical constituents are appropriately related. The paper [10] makes the specification that the s symbols do not appear in the quantum circuit, so we ignore them, and only the n symbol annihilations matter. For these $(n^L n) \to 1$ and $(nn^R) \to 1$ reductions we use a CNOT followed by a Hadamard gate, and then measure the final state. If the measurement result is $|00\rangle$, this indicates that the two "nounlike" vectors were appropriately correlated in meaning. Once all the qubits have been measured, if the final result is all zeroes, then all the meaning vectors match appropriately, and we predict that the sentence is true. Otherwise, something went wrong or there were inconsistent meanings, and we predict that the sentence is false. In practice, we can run the circuit many times, and count the proportion of results where the output is all zeroes, and use this ratio as an estimate of the probability that the sentence is true.

4.2.3 Parameter Optimization

The only remaining step is to find what parameters our word initialization gates should use. This corresponds to learning the word embeddings, as it is determining what vector the qubits corresponding to each word will be initialized as. To do this, we need a training data set, consisting of a vocabulary and several sentences with assignments to `True` (1) and `False` (0) values. Then we can define a loss function as the sum of squared deviations between predicted and actual values:

$$L(\vec{\theta}) = \sum_{\sigma} (l_{\sigma}^{\text{pr}}(\vec{\theta}) - l_{\sigma})^2,$$

where we sum over all sentences σ, l_σ is the actual truth value, and $l_\sigma^{\text{pr}}(\vec{\theta})$ is the predicted truth proportion using the quantum circuit parameterized with $\vec{\theta}$.

Since we don't have a closed-form solution for the gradient of this loss function with respect to the parameters $\vec{\theta}$, we require a gradient-free method. We use `minimizeSPSA` from the `noisyopt` Python package which perturbs the values of $\vec{\theta}$ in a random direction and then compares estimates of the loss function for both possible parameterizations, to estimate the gradient and move toward lower loss. This algorithm can yield decent accuracy even for small numbers of qubits; for more details on loss values, see my paper [1].

4.3 Information in Language

Shannon applied his theory of information to written language at several levels, including characters, words, and combinations of words [7]. The entropy of pairs of words appearing next to each other in text is less than twice that of individual words. Mathematically, since words are not independent, they share some information and context $H(X:Y) > 0$, so that the joint information is

$$H(X,Y) = H(X) + H(Y) - H(X:Y) < H(X) + H(Y).$$

This means that it is harder to understand text if you start in the middle, rather than having all of the previous context, because there is more uncertainty and information content associated with words taken out of context.

The idea of previous words giving us a good guess of the next word is used quite profitably in large language models (LLMs) today. As a slight oversimplification, LLMs can be said to repeatedly predict the most likely word to come next, with a very complicated approximation to the true distribution of word strings in a language. LLMs can give a convincing appearance of natural English text, with grammatical sentences that stay focused on a given topic. Even the model Shannon demonstrated, based on predicting one word given only the previous word, has some semblance of common English. Such models work very well for the aspects of language that can be reduced to flat relationships between two words in a text, but they seem to have much more trouble modeling deeper relationships combining several concepts, such as the meaning of a complicated sentence. LLMs can often be seen generating texts with reasonably matching words and structures, but completely contradictory meanings.

Compositional models such as DisCoCat aim to describe the recursive structure of language so that even deeper meanings can be correctly inferred. Instead of a shallow word-by-word approach, DisCoCat combines meanings in sensible chunks. This model also has implications for the structure of information in language. While classical models most often consider the distribution over n words in a row, or the conditional distribution for the next word, in compositional models of meaning it is most natural to consider the distribution of information between grammatically connected pieces of a sentence, such as the subject and the predicate. This would not work for classical n-word distributions, since a predicate may have varying numbers of words, and predicates of different length are incomparable, but in DisCoCat, all meaning vectors of a given type (like part of speech, or a particular combination, like predicate) are comparable, since they live in the same vector space. Even beyond a classical distribution over predicate vectors, DisCoCat allows us to interpret meaning vectors as quantum states, and if these vector spaces decompose as a product of subspaces, we can compute the von Neumann entropy for reduced density matrices corresponding to these subspaces. We explore this point further in Chapter 5.

Chapter 5

Connections

5.1 Word Vectors and Entanglement

How can we quantify the information content in language? One of the most interesting aspects of language is the correlation between different words, allowing us to view grammatically connected words as entangled in some way.

A standard measure of quantum entanglement is found in the CHSH inequality, as seen in Section 2.2. A recent paper [17] claims that language violates Tsirelson's bound by finding relative proportions of certain noun-verb pairs and plugging them into this CHSH inequality. Considering instances of the sentence "the animal acts", the following word pair frequencies were counted:

	growls	whinnies	snorts	meows
horse	0	464	202	0
bear	247	0	0	0
tiger	97	0	0	0
cat	41	0	0	331

Table 5.1 Counts of pairs of animals and actions [17]

To apply the CHSH inequality here, the authors must interpret this as two 2-state quantum systems. If each word is represented by its first letter, the first system "animal" has two different orthonormal bases $A = \{|h\rangle, |b\rangle\}$ and $A' = \{|t\rangle, |c\rangle\}$, and the second system "acts" has $B = \{|g\rangle, |w\rangle\}$ and $B' = \{|s\rangle, |m\rangle\}$. Comparing the results with the CHSH inequality, they get

$$E(A',B') + E(A,B') + E(A',B) - E(A,B) = \frac{331}{331} + \frac{202}{202} + \frac{97-41}{97+41} - \frac{-464-247}{464+247} = 3 + \frac{56}{138}.$$

This exceeds 2, indicating non-classical behavior, but it even exceeds Tsirelson's bound of $2\sqrt{2}$, which the authors interpret as language having beyond-quantum entanglement. They then fit this into the framework of quantum mechanics by allowing entangled measurements of two qubits. Their generalization is quite interesting and may provide insight into correlations in language, but for the purpose of this thesis we will argue against their 2-state interpretation, and instead embed the words in a larger space.

If we think about the meanings of the different animals, it does not make sense to think of $\{|h\rangle, |b\rangle\}$ and $\{|t\rangle, |c\rangle\}$ as orthogonal sets spanning the same space. The most natural space to consider should represent nouns, or at least animals, and there is no reason every animal should be a combination of a tiger and a cat. In particular, a horse can hardly be represented as a linear combination of tiger and cat, so these two pairs can hardly be said to span the same space.

A more likely interpretation of the situation is that $\{|h\rangle, |b\rangle, |t\rangle, |c\rangle\}$ live in a vector space of dimension greater than two, and that they are linearly independent. They may likely be non-orthogonal, and they probably should live in a space of dimension greater than 4, but for simplicity of example, let us assume that $\{|h\rangle, |b\rangle, |t\rangle, |c\rangle\}$ forms an orthonormal basis for the state space of the "animal" system, and that $\{|g\rangle, |w\rangle, |s\rangle, |m\rangle\}$ does the same for the "acts" system. Given this, we could represent the "animal acts" state of Table 5.1 by

$$\frac{1}{\sqrt{1382}}\left(\sqrt{464}\,|hw\rangle+\sqrt{202}\,|hs\rangle+\sqrt{247}\,|bg\rangle+\sqrt{97}\,|tg\rangle+\sqrt{41}\,|cg\rangle+\sqrt{331}\,|cm\rangle\right). \qquad (5.1)$$

Notice that the relative phases are undetermined by the probabilities of co-occurrences. In fact, we need not even presuppose any entanglement or superposition here, since the same probability distribution could arise from a classical mixture of the word pair combinations, in the density matrix

$$\begin{aligned}\rho_{\text{classical}}=\frac{1}{1382}\Big(&464\,|hw\rangle\langle hw|+202\,|hs\rangle\langle hs|+247\,|bg\rangle\langle bg|\\&+97\,|tg\rangle\langle tg|+41\,|cg\rangle\langle cg|+331\,|cm\rangle\langle cm|\Big).\end{aligned} \qquad (5.2)$$

Measuring either of these states state would give a distribution similar to the one found in the table, and therefore shows how there is no need for beyond-quantum behavior in this interpretation. Counts from the measurement cannot be inserted into the CHSH inequality, since the inequality only applies to pairs of 2-state quantum systems, each with two different orthogonal measurement bases. For example, it is incorrect to treat $A=\{|h\rangle,|b\rangle\}$ and $B=\{|g\rangle,|w\rangle\}$ as complete orthogonal bases in order to compute

$$E(A,B)=\frac{(\langle h|-\langle b|)\rho_{\text{classical}}(|g\rangle-|w\rangle)}{(\langle h|+\langle b|)\rho_{\text{classical}}(|g\rangle+|w\rangle)}=\frac{-464-247}{464+247}=-1$$

to plug into the CHSH inequality. In fact, since we have found a quantum state which yields the measured probability distribution, it will necessarily obey any laws that apply to quantum systems.

5.2 Von Neumann Entropy and Language

We have seen that correlations in language can be modeled by quantum states or even just by classical probability distributions over word counts. If we want to consider the combination "animal acts" as complete and unambiguous, while the meaning of its individual words "animal" and "acts" are ambiguous, then we can do this by considering "animal acts" to be an entangled pure state, as in the expression (5.1). Since the entire system is described by a pure state, its von Neumann entropy is zero, but we can compute a nonzero von Neumann entropy for its subsystems, the individual words. The word "animal" has a subsystem density matrix

$$\begin{aligned}\rho_{\text{Animal}} = \frac{1}{1382}\Big(&666\,|h\rangle\langle h| + 247\,|b\rangle\langle b| + 97\,|t\rangle\langle t| + 372\,|c\rangle\langle c|\\ &+\sqrt{247\cdot 97}(|b\rangle\langle t| + |t\rangle\langle b|) + \sqrt{247\cdot 41}(|b\rangle\langle c| + |c\rangle\langle b|) + \sqrt{97\cdot 41}(|t\rangle\langle c| + |c\rangle\langle t|)\Big)\end{aligned}$$

$$\approx \begin{array}{c} \\ \text{horse} \\ \text{bear} \\ \text{tiger} \\ \text{cat} \end{array} \begin{array}{c} \begin{array}{cccc} \text{horse} & \text{bear} & \text{tiger} & \text{cat} \end{array} \\ \begin{bmatrix} 0.482 & 0 & 0 & 0 \\ 0 & 0.179 & 0.112 & 0.073 \\ 0 & 0.112 & 0.071 & 0.046 \\ 0 & 0.073 & 0.046 & 0.269 \end{bmatrix} \end{array},$$

which has eigenvectors

$$\left\{\begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0.56301\\0.352821\\0.747353\end{bmatrix}, \begin{bmatrix}0\\-0.633279\\-0.396856\\0.664427\end{bmatrix}, \begin{bmatrix}0\\-0.531015\\0.847363\\0\end{bmatrix}\right\}$$

with eigenvalues $\{0.48191, 0.345574, 0.172516, 0\}$, for a total entropy of 1.47463 bits. The word "acts" is represented by the density matrix

$$\begin{aligned}\rho_{\text{Acts}} = \frac{1}{1382}\Big(&464\,|w\rangle\langle w| + 202\,|s\rangle\langle s| + 385\,|g\rangle\langle g| + 331\,|m\rangle\langle m|\\ &+\sqrt{464\cdot 202}(|w\rangle\langle s| + |s\rangle\langle w|) + \sqrt{41\cdot 331}(|g\rangle\langle m| + |m\rangle\langle g|)\Big)\end{aligned}$$

$$\approx \begin{array}{c} \\ \text{growls} \\ \text{whinnies} \\ \text{snorts} \\ \text{meows} \end{array} \begin{array}{c} \begin{array}{cccc} \text{growls} & \text{whinnies} & \text{snorts} & \text{meows} \end{array} \\ \begin{bmatrix} 0.336 & 0.222 & 0 & 0 \\ 0.222 & 0.146 & 0 & 0 \\ 0 & 0 & 0.279 & 0.084 \\ 0 & 0 & 0.084 & 0.240 \end{bmatrix} \end{array}.$$

which has the different eigenvectors

$$\left\{\begin{bmatrix}0.834684\\0.55073\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\0.782875\\0.622179\end{bmatrix}, \begin{bmatrix}0\\0\\-0.622179\\0.782875\end{bmatrix}, \begin{bmatrix}-0.55073\\0.834684\\0\\0\end{bmatrix}\right\},$$

but the same eigenvalues $\{0.48191, 0.345574, 0.172516, 0\}$, yielding the exact same entropy of 1.47463 bits. The combination of these two subsystems to yield a pure state with zero entropy is consistent with the fact that each subsystem of a bipartite system in a pure state must have the same von Neumann entropy.

If we instead use the assumption of a completely classical mixture (5.2), then we get the same density matrices, just with no coherence terms (meaning that only the diagonal terms remain):

$$\rho_{\text{Animal}} = \begin{array}{c} \\ \text{horse} \\ \text{bear} \\ \text{tiger} \\ \text{cat} \end{array} \begin{array}{c} \begin{array}{cccc} \text{horse} & \text{bear} & \text{tiger} & \text{cat} \end{array} \\ \left[\begin{array}{cccc} 0.482 & 0 & 0 & 0 \\ 0 & 0.179 & 0 & 0 \\ 0 & 0 & 0.071 & 0 \\ 0 & 0 & 0 & 0.269 \end{array}\right] \end{array}$$

and

$$\rho_{\text{Acts}} = \begin{array}{c} \\ \text{growls} \\ \text{whinnies} \\ \text{snorts} \\ \text{meows} \end{array} \begin{array}{c} \begin{array}{cccc} \text{growls} & \text{whinnies} & \text{snorts} & \text{meows} \end{array} \\ \left[\begin{array}{cccc} 0.336 & 0 & 0 & 0 \\ 0 & 0.146 & 0 & 0 \\ 0 & 0 & 0.279 & 0 \\ 0 & 0 & 0 & 0.240 \end{array}\right] \end{array}.$$

The eigenvalues of these diagonal matrices are just the values appearing on the diagonals, from which we can compute that their entropies are 1.73017 and 1.94165, respectively. For comparison, the entropy of the composite classical mixture (5.2) is 2.29155. This agrees with the fact that for combinations of classical probability distributions, $\max(S(A), S(B)) \leq S(AB) \leq S(A) + S(B)$.

5.3 Conclusions

The two situations we have considered here, namely, the pure state and the classical mixture, do not exhaust the possible quantum density matrices consistent with the empirical word counts. In fact, there is an entire spectrum of possibilities for coherence, between the completely classical mixture and the pure state. In this general case, the whole sentence has an intermediate amount of entropy, and the individual words have unequal but similar entropies. Another aspect left unspecified when considering only word counts is the relative phase between the components of (5.1). Our choice to neglect phases simplifies the expressions and coincides with the procedure in Section 3 of Tai-Danae Bradley's thesis [18].

Regardless of the specifics, the analogy we have explored suggests some properties of the distribution of information content in language. If a sentence has ambiguities represented by a classical probability distribution, then the entropies of its subparts are relatively unconstrained, just needing to satisfy properties of joint Shannon entropy. However, if a simple sentence can be represented by a (nearly) pure quantum state, then if we divide it into two subparts, like subject and predicate, then the entropies of the two parts must be (nearly) equal.

The example sentence we considered, "the animal acts", only involves a narrow amount of ambiguity, namely between different animals, and between different animal actions. In such a restricted case, it may seem plausible that there is a significant relationship between all animals considered, and even coherence to the point that the sentence is almost pure. However, if we consider the distribution of subject and predicate meanings over all possible sentences, it seems likely that there will be much less coherence, but existing coherences may reveal which sentences are analogically related. It would be interesting to explore this distribution more quantitatively using the DisCoCat framework on a variety of sentences. It could even be insightful to explore univariate distributions, such as the distribution of predicate vectors, since this is only generally possible in a compositional framework like DisCoCat.

Bibliography

[1] T. Draper, "Question Answering on Quantum Computers," *Journal of the Utah Academy of Sciences, Arts, & Letters* **100** (Mar, 2024) 321–332. https://www.utahacademy.org/wp-content/uploads/2024/03/JUASAL-2023-full-text-final2.pdf.

[2] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, USA, 10th ed., 2011.

[3] J. S. Bell, "On the Einstein Podolsky Rosen paradox," *Physics Physique Fizika* **1** (Nov, 1964) 195–200. https://link.aps.org/doi/10.1103/PhysicsPhysiqueFizika.1.195.

[4] J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt, "Proposed Experiment to Test Local Hidden-Variable Theories," *Phys. Rev. Lett.* **23** (Oct, 1969) 880–884. https://link.aps.org/doi/10.1103/PhysRevLett.23.880.

[5] B. S. Cirel'son, "Quantum generalizations of Bell's inequality," *Letters in Mathematical Physics* **4** (1980) 93–100. http://www.tau.ac.il/~tsirel/download/qbell80.html.

[6] J. Preskill, "Lecture Notes for Ph219/CS219: Quantum Information and Computation Chapter 4," *California Institute of Technology* (2001) 22–24. http://theory.caltech.edu/~preskill/ph229/notes/chap4_01.pdf.

[7] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal* **27** no. 3, (1948) 379–423.

[8] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. 3rd ed., 2024. https://web.stanford.edu/~jurafsky/slp3/.

[9] B. Coecke, *From Quantum Foundations via Natural Language Meaning to a Theory of Everything*, pp. 63–80. Springer International Publishing, Cham, 2017. https://doi.org/10.1007/978-3-319-43669-2_4.

[10] K. Meichanetzidis, A. Toumi, G. de Felice, and B. Coecke, "Grammar-aware sentence classification on quantum computers," *Quantum Machine Intelligence* **5** no. 1, (Feb, 2023) . https://doi.org/10.1007/s42484-023-00097-1.

[11] J. Lambek, "The mathematics of sentence structure," *American Mathematics Monthly* **65** (1958) .

[12] D. Kartsaklis, I. Fan, R. Yeung, A. Pearson, R. Lorenz, A. Toumi, G. de Felice, K. Meichanetzidis, S. Clark, and B. Coecke, "lambeq: An Efficient High-Level Python Library for Quantum NLP," arXiv:2110.04236 [cs.CL].

[13] B. Coecke, M. Sadrzadeh, and S. Clark, "Mathematical foundations for a compositional distributional model of meaning," *Linguistic Analysis* **36** (2010) , arXiv:1003.4394 [cs.CL].

[14] B. Coecke and A. Kissinger, *Picturing Quantum Processes: A First Course in Quantum Theory and Diagrammatic Reasoning*. Cambridge University Press, 2017.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, eds., vol. 26. Curran Associates, Inc., 2013. https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

[16] Qiskit contributors, "Qiskit: An Open-source Framework for Quantum Computing,". https://doi.org/10.5281/zenodo.2573505.

[17] L. Beltran and S. Geriente, "Quantum Entanglement in Corpuses of Documents," *Foundations of Science* **24** no. 2, (Jun, 2019) 227–246. https://doi.org/10.1007/s10699-018-9570-2.

[18] T.-D. Bradley, "At the Interface of Algebra and Statistics," arXiv:2004.05631 [quant-ph]. PhD thesis, CUNY Graduate Center.

Index