Using Density Matrices in a Compositional Distributional Model of Meaning

572640

Kellogg College

University of Oxford

A thesis submitted for the degree of

Master of Science

Trinity 2014

This thesis is dedicated to
my parents Aslı and Babür Balkır
and my grandmothers Suna Ziler and Türkan Balkır

Acknowledgements

I would like to thank my supervisor Bob Coecke. He has been a great resource for ideas, and our meetings never failed to inspire different approaches to the problems I was trying to solve. I would also like to thank him and Edward Grefenstette for offering the reading course on compositional distributional models of meaning. Their patience and engagement made it the best possible setting to learn the background material for this dissertation.

I would like to thank Kevin Milner for never turning down a request to answer my questions over a pint, no matter how silly they were. He was a dear friend in Montreal, and I'm thankful for his friendship in Oxford and for his wisdom in all things quantum. I would also like to thank him and Martti Karvonen for proofreading this dissertation.

I would like to thank my undergraduate professor Prakash Panangaden for showing me some wonderfully interesting pieces of mathematics and computer science, being a great mentor, and trusting in my abilities more than I have. He is the primary reason I ended up doing this program.

I would like to thank everybody involved in MFoCS. Lastly, I would like to thank everybody that I have come to call my friend in the last year. One year was too short, but still you made Oxford feel like home.

Abstract

In this dissertation I present a framework for using density matrices instead of vectors in a distributional model of meaning. I present an asymmetric similarity measure between two density matrix representations based on relative entropy and show that this measure can be used for hyponymy-hypernymy relations. It is possible to compose density matrix representations of words to get a density matrix representation of a sentence. This map respects the generality of individual words in a sentence, taking sentences with more general words to more general sentence representations.

Contents

Chapter 1

Introduction

The term *distributional models of meaning* is almost synonymous with the term *vector space models of meaning*. This is because vector spaces are the simplest, most natural candidate for modeling contextual similarity between words. There are fast, efficient algorithms for obtaining vector representations and for processing them. Vector spaces are basic and widely used structures in a variety of disciplines, and using them as a basis for modeling semantic content provides a rich array of methods at linguists' disposal. Their obvious practicality, however, does not guarantee that they possess the expressive power needed to model meaning. The fact that current distributional models fall short of modeling subsumption and entailment is well noted [23], but most of the work has been focused on developing more sophisticated metrics on vector representations.

In this dissertation I suggest the use of density matrices instead of vector spaces as the basic distributional representations for the meanings of words. Density matrices are widely used in quantum mechanics, and are a generalization of vectors. The advantages of using density matrices to model meaning are as follows:

- Density matrices have the expressive power to represent all the information vectors can represent: they are a suitable implementation of the distributional hypothesis.
- They come equipped with a measure of information content, and so provide a natural way of implementing asymmetric relations between words such as hyponymy-hypernymy relations.
- They form a compact closed category. This allows the previous work on obtaining representations for meanings of sentences from the meaning of words applicable to density matrices without any major modifications.

- Density matrix representations of subsumption relations between words work well with the categorical map from meanings of words to the meaning of the sentence. This promises, from representations of individual words, a method to obtain entailment relations on the level of sentences.

This work is organized as follows: in Chapter 2, I present two main paradigms of computational linguistics, formal semantics and distributional semantics. I review the relevant work done to address the shortcomings of distributional semantics on modeling compositionality and entailment. I mention some literature that uses similar methods with this work, such as information theoretic measures and density matrices. I also review some relevant distance measures on semantic networks. These models represent meaning based on asymmetric relations as much as symmetric ones, and provide a natural venue for investigating and comparing measures for hyponymy.

Chapter 3 introduces the linear algebraic notation I use throughout the dissertation, borrowed from quantum mechanics. Chapter 4 presents the relevant measures to quantify information content of discrete probability distributions and their generalizations to density matrices. Chapter 5 introduces the categorical framework that the compositional distributional model presented in Chapter 6 uses. Chapter 7 shows that density matrices fit into the same categorical framework.

Chapter 8 is the culmination of all the previous material, where I explain what kind of a distributional model would use the full power of the density matrix formalism, review three candidates for a similarity measure, and suggest an asymmetric representativeness measure based on relative entropy together with a partial order that it imposes on density matrices. I show that according to the representativeness measure, the composition map presented in chapter 6 takes more general words to more general sentences, irrespective of the choice for the sentence space. Chapter 9 illustrates the ideas laid out in chapter 8 by applying them to example sentences.

This dissertation presents theoretical work that demonstrates using density matrices instead of vector spaces for modeling meaning provides promising methods to overcome some basic shortcomings of the current state-of-the-art distributional models. I suggest density matrices as a mathematical framework, and leave the implementation and testing of these ideas for further work.

Chapter 2

Literature Review

2.1 Formal semantics

There are two main paradigms for computational natural language semantics. First is formal semantics, which takes a model-theoretic approach to modeling meaning. In formal semantics, the meaning of a sentence is a mapping from the sentence to models in which it is deemed to be true. This approach was introduced mainly by Richard Montague and is often referred to as *Montague grammar*. The spirit of his method is well reflected in the name of his famous paper *"English as a Formal Language"* in which Montague declares: "I reject the contention that an important theoretical difference exists between formal and natural languages." [28, p.189]. He made use of a typed intensional logic together with a model structure that included a set of *possible worlds*. He defined the denotation of a sentence as a function from possible worlds and moments of time to truth values. His program was based on converting sentences of natural language into logic formulas, which were then assigned a semantic interpretation in the form of a denotation. See [33, Lecture 2, Appendix] for a more detailed presentation of his intensional logic.

Formal semantics succeeded in providing an account of two important aspects of natural language semantics:

1. **Compositionality**: This is the idea that the meaning of the whole is a function of its parts, and the way they are syntactically combined [31]. One of the primary goals of formal semantics is to lay out the precise way the syntax of a sentence affects its meaning, since a satisfactory linguistic theory must address how a competent speaker can understand the meanings of novel sentences [32].

   One of the most important innovations Montague brought into linguistic semantics from the field of formal logic was lambda terms. He used lambda

abstractions to model the meanings of some words as functions, and function application served as the composition operation that relates the meanings of individual words and returns the meaning of the larger phrase.

2. **Entailment**: In Montague's own words, the aim of semantics should be to "characterize the notion of a true sentence (under a given interpretation) and of entailment" [29, p.223]. The model theoretic basis of formal semantics provides the necessary framework for such characterization. The denotation of a sentence characterizes in which worlds a sentence is true, and entailment spans all possible worlds: sentence A is defined to entail sentence B if in all models in which the interpretation of A is true, the interpretation of B is also true[18].

2.2 Distributional semantics

Distributional semantics adopt a corpus based approach to natural language semantics. The theoretical assumption behind a wide range of methods that fall under distributional semantics can be summarized by Firth's dictum: "you shall know the word by the company it keeps"[9]. The idea that contexts are intimately connected with meaning can be traced back to Harris [15, p.156]:

> If we consider words of morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference in meaning correlates with difference of distribution.

Distributional methods rely on co-occurrence statistics of words to represent the words' meanings as vectors in a high dimensional vector space. Consequently, distributional models are also referred to as *vector space* or *semantic space* models. Vector spaces provide a truly distributional representation: the semantic content of a word is defined in relation to the words it is close to and far from in meaning. This aspect gives rise to a model of meaning that is inherently gradual.

The distributional representations are usually obtained from large corpora. The simplest way to obtain such a representation of a word w is to pick a set of context words, then count the number of times w occurs in proximity with each of the context words. Then the distributional representation of w would be a normalized vector where each context word c corresponds to a basis vector $\overrightarrow{b_c}$, and the co-occurrence

frequency of w and c becomes the weight of $\overrightarrow{b_c}$ in the vector representation of w. There are a number of computational methods that improve on this basic idea, such as *Latent Semantic Analysis (LSA)* [6] and *Hyperspace Analogue to Language (HAL)* [25].

Two main properties of the distributional models make them particularly suitable for various Natural Language Processing applications:

- It is possible to automatically obtain meaning representations of words from text.
- The geometric and gradual representations provide several natural similarity measures on the representations.

Success stories of the applications of distributional semantics include automatic thesaurus extraction [14, 5], automated essay marking [40], and word-sense discrimination [27, 36].

The weaknesses of distributional semantics are exactly the strengths of formal semantics: compositionality, inference and logic. I will focus on the previous work done to address the first two of these.

Compositionality in distributional semantics. Distributional semantics operates almost entirely on the word level, since its methods cannot be applied to entire sentences. On a practical level this is because entire sentences occur too infrequently to obtain meaningful statistics. On a more conceptual level, trying to assign meanings to sentences by treating them as single blocks would fail as a theory of meaning, since it would ignore the fundamental fact that humans understand sentences they've never heard before.

There has been a recent interest in methods of composition within the distributional semantics framework. There are a number of composition methods in literature. See [12] for a survey of compositional distributional models and a discussion of their strengths and weaknesses. The work in this dissertation is built upon [4], a compositional model based on category theory. This model was shown to outperform the competing compositional models in [13].

Entailment in distributional semantics. The work on entailment in distributional semantics has been based almost entirely on lexical entailment, where it roughly translates as follows: w entails v if the meaning of a word w is included in the meaning

of a word v, or equivalently if w *is-a* v. I will refer to this property as the *subsumption relation* between v and w.

The notion of distributional similarity is a somewhat loose notion that does not necessarily capture whether the meaning will be preserved when a word is replaced with another one, as noted in [11]. The usual similarity measure falls short on this task since it is symmetric, whereas whether one can use a given word in place of another is clearly not. One can freely say *furniture* instead of *table*, but not the other way around. This distinction is important for a number of Natural Language Processing applications such as Question Answering, Information Retrieval, Information Extraction, and Text Categorization [19].

A number of non-symmetric similarity measures [43, 3, 19, 23] rely on some variation of the *Distributional Inclusion Hypothesis*.

> **Distributional Inclusion Hypothesis**[19]: If u is semantically narrower than v, then a significant number of salient distributional features of u are also included in the feature vector of v.

[11] break the Distributional Inclusion Hypothesis in two:

- **Hypothesis 1:** If $v => w$ then all the characteristic features of v is expected to appear in w.
- **Hypothesis 2:** If all the characteristic features of v appear in w, then $v => w$.

They develop an *Inclusion Testing Algorithm* to test the two hypotheses, and conclude that while the first one is verified by their experiments, the second hypothesis does not fully hold. One of their suggestions for increasing the prediction power of their method is to include more than one word in the features.

A similar idea is *confusion probability* [38, 22], which measures the degree to which a word w can be substituted in the contexts v appears in. *α-skew divergence*, suggested in [22], is another asymmetrical similarity measure. It is based on Kullback-Leibler divergence (KL-divergence) , and has been shown to outperform the other symmetric measures considered in the study.

[35] use a measure based on entropy to detect hyponym-hypernym relationships in given pairs. The measure they suggest rely on the hypothesis that hypernyms are semantically more general than hyponyms, and therefore tend to occur in less informative contexts. [16] rely on a very similar idea, and use KL-divergence between the target word and the basis words to quantify the semantic content of the target

word. They conclude that this method performs equally well in detecting hyponym-hypernym pairs as the overall frequency of the word in corpus, and reject the hypothesis that more general words occur in less informative contexts. Their method differs from mine in that they use relative entropy to quantify the overall information content of a word, and not to compare two target words to each other.

[1] combines entailment and composition to some degree. Their method goes beyond lexical entailment and applies to some types of noun-phrases. They build on the intuition that an adjective noun pair such as *red car* almost always entails the noun *car*. They use the distributional data on such adjective-noun pairs to train a classifier, which is then utilized to detect novel noun pairs that have the same relation, such as *dog* and *animal*. They use the same method on detecting entailment in quantifier-noun pairs. The classifier trained on pairs such as *many dogs - some dogs* successfully predict the entailment relation between *all cats - several cats*.

From meaning as a point to meaning as a region. From cognitive science, Gärdenfors suggests a model of conceptual representation that he calls *conceptual spaces* [10]. This model treats concepts as regions in a space where the bases are perceptual dimensions, and natural properties are those that occupy a convex region of the space. Going from points to regions has the obvious application of characterizing the generality of meanings and subsumption relations. [7] implements these ideas in a co-occurrence space to generalize the representation of a word from a vector to a region. She concludes that this approach is successful in detecting hyponymy-hypernymy relations.

Density matrices in linguistics and IR. [2] use density matrices to model context effects in a conceptual space. In their quantum mechanics inspired model, words are represented by mixed states and each eigenstate represents a sense of the word. Context effects are then modeled as quantum collapse.

The most notable suggestion of using density matrices in a related area is [41], where a theory of Information Retrieval based on density matrices is sketched out. [41] connects the logic of the space and density matrices as follows:

Let $\mathbf{E}$ be a projector in a Hilbert space. Then the order relation

$$\mathbf{E} \leq \mathbf{F} \text{ if and only if } \mathbf{FE} = \mathbf{E}$$

makes the set of projectors in a Hilbert space into a complete lattice, and hence $\leq$ provides an entailment relation.

This logical structure is tied to density matrices conceptually by pointing out that a density matrix is a generalized question, thus can be decomposed by the Spectral Theorem into a linear combination of yes-or-no questions. Yes-or-no questions correspond to projectors since any projector has exactly two eigenvalues, 1 and 0.

2.3 Semantic network models

A semantic network is a network representation of meaning, with nodes representing words or classes of words, and directed and labeled edges representing certain relations that hold between the classes. Most widely used semantic network is WordNet [8], a hand coded lexical database for English that contains more than 118,000 word forms and 90,000 word senses. The words are organized into groups of synonyms called *synsets*. Wordnet defines a number of non-reflexive relations between synsets, as well as the reflexive relations *synonymy* and *antonymy*.

- **Hyponymy** is the *is a* relationship between nouns: *dog* is a hyponym of *animal*. The opposite relation is called **hypernymy**: *furniture* is a hypernym of *table*. Because there is mostly a single hypernym for many hyponyms, this relation organizes nodes into a hierarchical tree structure. The tree is sometimes referred to as an *is-a tree*.

- **Meronymy** is the whole-part relationship between nouns: *wheel* is a meronym of *car*.

- **Troponymy** is the verb equivalent of hyponymy: *stroll* is a troponymy of *walk*.

- **Entailment** is a relationship between verbs: *snore* entails *sleep*.

The hierarchical structure of hyponymy relations are suggestive of information theoretic interpretations: the higher up an *is-a* tree one goes, less specific the words become, and this intuitively correlates with less information. Indeed, there has been several applications of these ideas to semantic networks:

Applications of entropy in semantic networks. [34] offers a model to evaluate semantic similarity in a taxonomy based on the notion of information content. His starting point is the hypothesis that "the similarity of two concepts is the extent to which they share information". He assumes a taxonomical representation, and defines the similarity of two concepts c_1 and c_2 as the entropy of their lowest common ancestor. [24] also define a similarity measure based on information content that

they derive from reasonable assumptions for similarity. They demonstrate that it is applicable to a number of representations of meaning, including a semantic network.

Chapter 3

Pure States and Mixed States in Quantum Mechanics

Quantum mechanics is formulated using complex Hilbert spaces. Hilbert spaces are real or complex complete inner product spaces. Hilbert spaces used in linguistics are finite dimensional, thus trivially satisfy the completeness condition. A real vector space is a Hilbert space if it is equipped with an inner product. This qualifies meaning spaces in distributional semantics as Hilbert spaces, so the mathematical notation developed to reason about quantum mechanics can be applied to meaning spaces as well. I will use this notation, referred to as the *Dirac notation* for the rest of my dissertation. Here I introduce the notation together with its terminology from quantum mechanics, following the presentation in [30] closely. What these linear algebraic concepts correspond to in quantum mechanics is not of primary relevance.

Pure states. In quantum mechanics, pure states are unit vectors in a Hilbert space. The standard notation for a pure state is the following:

$$|\psi\rangle$$

Where ψ is a label and $|\cdot\rangle$ indicates that it is a vector (or equivalently, a pure state). This is referred to as a *ket*. $|\psi\rangle$ can also be seen as an operator $\mathbb{F} \to V$, where $\mathbb{F}$ is the underlying field. Notice that such operators are in one-to-one correspondence with the elements of a Hilbert space. The dual of a vector $|\psi\rangle$ is the *effect* corresponding to the state, and it is the dual operator $V \to \mathbb{F}$. It is also referred to as a *bra*, and it is written as the following:

$$\langle\psi|$$

This the Hermitian conjugate of $|\psi\rangle$.

The inner product of two vectors $|\psi\rangle$ and $|\phi\rangle$ is written as $\langle\psi|\phi\rangle$, and as the notation suggests, is calculated by multiplying the Hermitian conjugate of $|\psi\rangle$ with $|\phi\rangle$. $\langle\psi|\phi\rangle = \langle\phi|\psi\rangle$, so it doesn't matter whether the conjugation is applied to the first or the second vector.

Outer product representation. A useful way to represent linear operators is called the *outer product* representation. Let $|v\rangle$ be a vector in a Hilbert space V and $|w\rangle$ be a vector in a Hilbert space W. $|w\rangle\langle v|$ is defined to a the linear operator from V to W, and its action is defined by:

$$(|w\rangle\langle v|)(|v'\rangle) \equiv |w\rangle\langle v|v'\rangle = \langle v|v'\rangle|w\rangle$$

So the operator $|w\rangle\langle v|$ acts on $|v'\rangle \in V$ by taking it to $\langle v|v'\rangle|w\rangle$, which is the same operation as multiplying $|w\rangle \in W$ with the scalar $\langle v|v'\rangle$.

An arbitrary operator $A : V \to W$ can be written out using the outer product formulation. Let $\{|v_i\rangle\}_i$ and $\{|w_j\rangle\}_j$ be bases for V and W respectively, and $[m_{ij}]$ be the matrix representation of the operator A in these chosen bases. Then,

$$A = \sum_{i,j} m_{ij}|w_j\rangle\langle v_i| \quad \text{and}$$
$$m_{ij} = \langle w_j|A|v_i\rangle$$

Eigenvectors and eigenvalues. A *diagonal representation* of an operator A on a vector space V is:

$$A = \sum_i \lambda_i|i\rangle\langle i|$$

Vectors $|i\rangle$ make up an orthonormal basis A diagonalizes in, hence they are eigenvectors of A. λ_i are the values in the diagonal entries: the eigenvalues corresponding to each $|i\rangle$. With the Dirac notation it is easy to check that $A|i\rangle = \lambda_i|i\rangle$.

Projectors. An operator that is equal to its own adjoint is called *Hermitian* or *self-adjoint*. Projectors are a class of Hermitian operators that are in one-to-one correspondence with the subspaces of a Hilbert space. These are the operators that are defined to act as the identity morphism on a subspace V, and as the zero morphism on $V^\perp$, the orthogonal complement of V.

Definition 3.1. Suppose W is a k-dimensional subspace of the d-dimensional Hilbert space V. It is possible to construct an orthonormal basis $|1\rangle, \ldots, |d\rangle$ for v such that $|1\rangle, \ldots, |k\rangle$ is an orthonormal basis for W using the *Gram-Schmidt procedure* [30, p.66].Then, the **projector** onto the subspace W is:

$$P \equiv \sum_{i=1}^{k} |i\rangle\langle i|$$

Equivalently, P is a projector if it satisfies $P = P^2$.

Theorem 3.2. *(**Real Spectral Theorem**) An operator M on a real vector space V is hermitian if and only if it is diagonal with respect to some orthonormal basis for V.*

For the proof, see [30, p.72]

Mixed states. A generalization of the idea of pure states are mixed states. Mixed states are defined to be probability distributions over ensembles of pure states, or equivalently as probability distributions on subspaces of a Hilbert space. The mathematical tool used to express this concept is called the *density operator*:

Definition 3.3. Given a set $\{p_i, |\phi\rangle_i\}_i$ where $\{|\phi\rangle_i\}$ is a set of orthonormal pure states and $\{p_i\}$ is a probability distribution over them, the corresponding **density operator** or **density matrix** is:

$$\rho \equiv \sum_i p_i |\phi_i\rangle\langle\phi_i|$$

Definition 3.4. A **positive operator** is an operator such that for any vector $|v\rangle$, $\langle v|A|v\rangle$ is a real, non-negative number.

Theorem 3.5. *An operator ρ is a density operator if and only if it is a positive Hermitian operator with trace equal to one.*

Proof. Any positive Hermitian operator has a diagonal representation $\sum_i \lambda_i |i\rangle\langle i|$, with non-negative eigenvalues λ_i by spectral decomposition. A positive Hermitian operator with trace equal to one would hence have $\sum_i \lambda_i = 1$, so it is a density operator for the ensemble $\{\lambda_i, |i\rangle\}_i$.

Any density operator $\rho = \sum_i p_i |i\rangle\langle i|$ has trace equal to one, since $\sum_i p_i = 1$. It is positive since for any vector $|v\rangle$:

$$\langle v|\rho|v\rangle = \sum_i p_i \langle v|i\rangle\langle i|v\rangle \tag{3.1}$$

$$= \sum_i p_i |\langle v|i\rangle|^2 \geq 0 \tag{3.2}$$

and it is clearly symmetric, so ρ is a positive Hermitian operator with trace equal to one. □

As for vector spaces, there is an inner product defined on density matrices:

Definition 3.6. If A and B are density matrices of same dimensions, the **trace inner product** of A and B is defined to be

$$\mathrm{tr}(A^T B)$$

If $A = \sum_i p_i |i\rangle\langle i|$ and $B = \sum_j q_j |j\rangle\langle j|$

$$\mathrm{tr}(A^T B) = \mathrm{tr}\left(\sum_i p_i |i\rangle\langle i| \sum_j q_j |j\rangle\langle j|\right) \tag{3.3}$$

$$= \sum_{i,j} p_i q_j \langle i|j\rangle \mathrm{tr}(|i\rangle\langle j|) \tag{3.4}$$

$$= \sum_{i,j} p_i q_j \langle i|j\rangle\langle i|j\rangle \tag{3.5}$$

$$= \sum_{i,j} p_i q_j \langle i|j\rangle^2 \tag{3.6}$$

Concretely, the trace inner product of two density matrices is the sum of their entry-wise products.

Trace inner product satisfies the conditions for an inner product:

Conditions for an inner product [30, p. 65] A function $(\cdot,\cdot) : V \times V \to \mathbb{F}$ is an inner product if:

- $(|v\rangle, |v\rangle) \leq 0$ with equality if and only if $|v\rangle = 0$.
- $(|v\rangle, |w\rangle) = (|w\rangle, |v\rangle)^*$
- $(\cdot,\cdot)$ is linear in the second argument:

$$\left(|v\rangle, \sum_i \lambda_i |w_i\rangle\right) = \sum_i \lambda_i (|v\rangle, |w\rangle)$$

One can check that the trace inner product indeed satisfies these properties, but an easier way to see this is to note that the space of n by n matrices is isomorphic to an n^2 dimensional vector space, and the trace inner product of two n by n density matrices is equal to the usual inner product of the n^2 dimensional vector counterparts.

Chapter 4

Classical and Quantum Information Measures

In this chapter I will introduce the origins and the ideas behind the information measures I apply to density matrix representations of meaning. These come from the field of information theory, and are based on the idea of entropy. I first introduce classical entropy and classical relative entropy, which are defined on discrete probability distributions. Then I present von Neumann entropy and quantum relative entropy. These generalize their classical counterparts to density matrices, and include classical entropy measures as special cases. This chapters follows [30].

Classical entropy. Information theory quantifies the amount of information one gains on average when an event E occurs in a probabilistic experiment, or equivalently, the uncertainty one has before any events take place.

If the probability of an event E occurring is already 1, one gains no information from observing E. The event was expected with certainty anyways. On the other hand if E has a very small probability of occurring, one would be quite surprised to observe it. The information one gains from the observation is high. If E and F are two independent events, the information one gain from observing both E and F is the sum of the information one gain from observing E and F separately:

$$I(EF) = I(E) + I(F)$$

Where $I(E)$ denotes the information of the event E.

If $I(E)$ is required to be a smooth function only dependent on the probability of E, then any I that suits the above conditions is necessarily defined as:

$$I(E) = k \log p_E$$

Where k is an arbitrary constant, and p_E is the probability of event E. Since the constant k does not affect the desired properties, it is assigned the value -1.

Definition 4.1. The **Shannon entropy** associated with a discrete probability distribution is its weighted average of information:

$$H(X) \equiv H(p_1, p_2, \ldots, p_n) \equiv -\sum_x p_x \log p_x$$

However, note that even if this is an intuitive justification for the definition of entropy, entropy is originally defined to quantify the minimal resources to store information. When the log in the definition is taken to be base 2, the formula gives the minimum number of bits needed to store the outcome of string of independent, identically distributed random variables without loosing any information.

Relative entropy. The idea of *relative entropy* is first introduced by Kullback and Leibler [20] and is therefore also referred to as *Kullback-Leibler divergence*.

Definition 4.2. Relative entropy is an entropy-like measure of the closeness of two probability distributions, $p(x)$ and $q(x)$, over the same index set, x:

$$H(p(x)||q(x)) \equiv -\sum_x p(x) \log \frac{p(x)}{q(x)} \equiv -H(X) - \sum_x p(x) \log q(x)$$

Here we use the convention that $0 \log 0 = 0$ and $x \log 0 = -\infty$ for $x > 0$.

Consider the case where the probability distribution is $p(x)$, but we erroneously believe that it is $q(x)$. The average information according to this erroneous belief is:

$$-\sum_x p(x) \log q(x)$$

The real amount of information we gain, however, is the regular Shannon entropy: $-\sum_x p(x) \log p(x)$. Relative entropy is the difference of these two. For the purposes of this dissertation, it is important that relative entropy is non-symmetric, and non-negative.

Theorem 4.3. *The relative entropy is non-negative,* $H(p(x)||q(x)) \geq 0$ *with equality if and only if* $p(x) = q(x)$ *for all* x.

Proof. $\log x \ln 2 + \ln x \leq x - 1$, for all positive x, with equality if and only if $x = 1$. So $-\log x \geq (1 - x)/\ln 2$.

$$H(p(x)||q(x)) = -\sum_x p(x) \log \frac{p(x)}{q(x)} \tag{4.1}$$

$$\geq \frac{1}{\ln 2} \sum_x p(x) \left(1 - \frac{q(x)}{p(x)}\right) \tag{4.2}$$

$$= \frac{1}{\ln 2} \sum_x (p(x) - q(x)) \tag{4.3}$$

$$= \frac{1}{\ln 2}(1 - 1) = 0 \tag{4.4}$$

Equality occurs on the second line only when $q(x)/p(x) = 1$, or in other words, if $q(x)$ and $p(x)$ are identical for all x. □

[42] explains the intuition behind the non-symmetry of relative entropy with the following analogy: suppose that we are given a coin that can either be completely fair, or always lands heads. If we toss the coin and it lands tails, we immediately know for sure that it is not completely unfair. However, if we keep tossing it and it keeps landing heads, even though we can be more and more certain that it is the completely unfair coin, we cannot be totally sure.

Quantum entropy. Quantum entropy, or *von Neumann entropy* is the generalization of Shannon entropy from discrete probability distributions to probability distributions on Hilbert spaces.

Definition 4.4. Von Neumann entropy of a density matrix ρ is defined as:

$$S(\rho) \equiv -tr(\rho \log \rho)$$

or equivalently as

$$s(\rho) \equiv -\sum_x \lambda_x \log \lambda_x$$

where λ_x are the eigenvalues of ρ.

Von Neumann entropy is always non-negative, and is zero only when the state is pure, or equivalently when it is a projector onto a one dimensional subspace. It is at most $\log d$ where d is the dimension of the Hilbert space. This happens when the state is completely mixed: I/d. This means that the density matrix assigns uniform probability to the entire space.

Quantum relative entropy. Quantum relative entropy is a generalization of classical relative entropy, and like its classical counterpart, offers a measure on the distinguishability of two density matrices ρ and σ. It was first considered by Umegaki [39].

Definition 4.5. The **(quantum) relative entropy** of two density matrices ρ and σ is:

$$N(\rho||\sigma) \equiv \operatorname{tr}(\rho \log \rho) - \operatorname{tr}(\rho \log \sigma)$$

where $0 \log 0 = 0$ and $x \log 0 = \infty$ when $x > 0$ by convention.

Proposition 4.6. *quantum relative entropy is always non-negative:*

$$N(\rho||\sigma) \geq 0$$

with equality if and only if $\rho = \sigma$.

Proof. Let $\rho = \sum_i p_i |i\rangle\langle i|$ and $\sigma = \sum_j q_j |j\rangle\langle j|$, where $\{|i\rangle\}_i$ and $\{|j\rangle\}_j$ are the orthonormal decompositions for ρ and σ respectively. Then,

$$\begin{aligned}
S(\rho||\sigma) &= \sum_i p_i \log p_i - \sum_i \langle i|\rho \log \sigma|i\rangle \\
&= \sum_i p_i \log p_i - \sum_i p_i \langle i| \log \sigma|i\rangle \\
&= \sum_i p_i \left(\log p_i - \langle i| \left(\sum_j \log(q_j)|j\rangle\langle j| \right) |i\rangle \right) \\
&= \sum_i p_i \left(\log p_i - \sum_j \langle i|j\rangle^2 \log q_j \right)
\end{aligned}$$

$[P_{ij}]$ where $P_{ij} \equiv \langle i|j\rangle^2$ is a doubly stochastic matrix, which means $\sum_i P_{ij} = \sum_j P_{ij} = 1$. By this property, it follows that $\sum_j P_{ij} \log q_j \leq \log(\sum_j P_{ij} q_j)$ with equality if and only if there exists a j for which $P_{ij} = 1$. Thus

$$S(\rho||\sigma) \geq \sum_i p_i \log \frac{p_i}{\sum_j P_{ij} q_j}$$

with equality if and only if $[P_{ij}]$ is a permutation matrix, that is, each row and each column of $[P_{ij}]$ has one 1 and all other entries are 0. From theorem 4.3 $\sum_i p_i \log \frac{p_i}{\sum_j P_{ij} q_j} \geq 0$, so $S(\rho||\sigma) \geq 0$. This means that ρ and σ are diagonal in the same basis and has the same eigenvalues for each eigenvector, so if $S(\rho||\sigma) = 0$ then $\rho = \sigma$. □

Proposition 4.7. *For two density matrices ρ and σ, $N(\rho||\sigma) = \infty$ if $supp(\rho) \cap \ker(\sigma) \neq 0$, and is finite otherwise.*

Proof. $N(\rho||\sigma) = \infty$ if and only if $\mathrm{tr}(\rho \log \sigma) = -\infty$, and is finite otherwise. Let $\sigma = \sum_i p_i |i\rangle\langle i|$.

$$\mathrm{tr}(\rho \log \sigma) = \sum_j \langle j|\rho \left(\sum_i \log p_i |i\rangle\langle i| \right) |j\rangle$$

Where $\{|j\rangle\} = \{|i\rangle\}$. If $i \neq j$, then $\log p_i |i\rangle\langle i|j\rangle = 0$, so the equation simplifies to:

$$\sum_j \langle j|\rho \log p_j |j\rangle = \sum_j \langle j|\rho|j\rangle \log p_j$$

For any j, $\langle j|\rho|j\rangle \log p_j = -\infty$ if and only if $p_j = 0$ and $\langle j|\rho|j\rangle \neq 0$, that is, $|j\rangle$ is both in the kernel of σ and the support of ρ. □

Chapter 5

Compact Closed Categories

Category theory has been developed as a general language for diverse mathematical structures in that it does not really concern itself with the internal components, but with the behaviour of the system. In other words, what is important from the perspective of category theory is not the objects but the morphisms. This point of view provides the flexibility for relating the structures used to represent grammatical types and the structures used to represent distributional meaning. These are pregroup algebras and vector spaces, and even though they are quite different, both are concrete instances of a compact closed category.

I will not define the notion of a category here, but refer the reader to [26].

5.1 Monoidal categories

Definition 5.1. [17] A **monoidal category C** is a category consisting of the following:

- a functor $\otimes\colon \mathbf{C} \times \mathbf{C} \to \mathbf{C}$ called the *tensor product*
- an object $I \in \mathbf{C}$ called the *unit object*
- a natural isomorphism whose components $(A \otimes B) \otimes C \xrightarrow{\alpha_{A,B,C}} A \otimes (B \otimes C)$ are called the *associators*
- a natural isomorphism whose components $I \otimes A \xrightarrow{\lambda_A} A$ are called the *left unitors*
- a natural isomorphism whose components $A \otimes I \xrightarrow{\rho_A} A$ are called the *right unitors*

These need to satisfy an additional *coherence* property which states that every well-formed equation built from $\circ$, $\otimes$, id , α, α^{-1}, λ, λ^{-1}, ρ and ρ^{-1} is satisfied.

Monoidal categories are used to model systems that have both sequantial and concurrent processes. The objects of the category are thought to be types of systems. A morphism $f : A \to B$ is a process that takes a system of type A to a system of type B. for $f : A \to B$ and $g : B \to C$, $g \circ f$ is the composite morphism that takes a system of type A into a system of type C by applying the process g after f. Morphisms of type $\psi : I \to A$ are called elements of A. The reason for that is the definition of a monoidal category does not include elements in the usual sense, for example as in $x \in A$ where A is a set. However, when the objects of a category does have elements, they are usually in bijective correspondence with the morphisms $\psi : I \to A$. This formalism therefore allows us to talk about elements without digressing from the terminology of category theory.

5.2 Compact closed categories

Definition 5.2. [4] A monoidal category is **compact closed** if for each object A, there are also left and right dual objects A^r and A^l, and morphisms

$$\eta^l : I \to A \otimes A^l \qquad \eta^r : I \to A^r \otimes A$$

$$\epsilon^l : A^l \otimes A \to I \qquad \epsilon^r : A \otimes A^r \to I$$

that satisfy the equations

$$(1_A \otimes \epsilon^l) \circ (\eta^l \otimes 1_A) = 1_A$$

$$(\epsilon^r \otimes 1_A) \circ (1_A \otimes \eta^r) = 1_A$$

$$(\epsilon^l \otimes 1_{A^l}) \circ (1_{A^l} \otimes \eta^l) = 1_{A^l}$$

$$(1_{A^r} \otimes \epsilon^r) \circ (\eta^r \otimes 1_{A^r}) = 1_{A^r}$$

The maps of compact categories are used to represent *correlations*, and in categorical quantum mechanics they model maximally entangled states. η and ϵ maps are useful in modeling the interactions of the different parts of a system. To see how this relates to natural language, consider a simple sentence with an object, a subject and a transitive verb. The meaning of the entire sentence is not simply an accumulation of the individual words, but depends on how the transitive verb relates the subject and the object. The η and ϵ maps provide the mathematical formalism to specify such interactions. The distinct left and right adjoints ensure that compact closed categories can take word order into account.

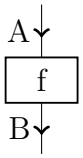
5.3 Graphical calculus

Graphical calculus for monoidal categories. There is a graphical calculus used to reason about monoidal categories which works through the following correspondence:

Theorem 5.3. *[4] An equational statement between morphisms in a monoidal category is provable from the axioms of monoidal categories if and only if it is derivable in the graphical language.*

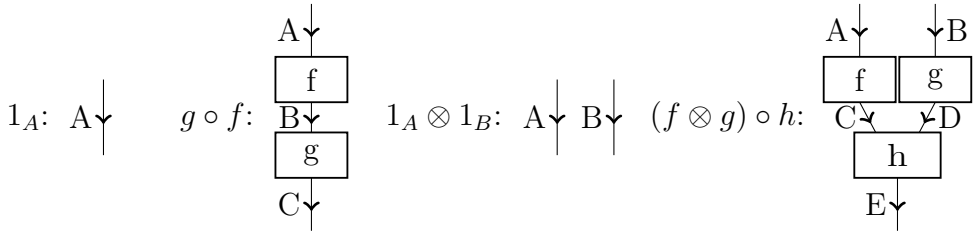
For an indepth presentation of the graphical calculus and the proof of the theorem see [37]

In the graphical language, objects are wires, and morphisms are boxes with incoming and outgoing wires of types corresponding to the input and output types of the morphism. The morphism $f : A \to B$ is depicted as:

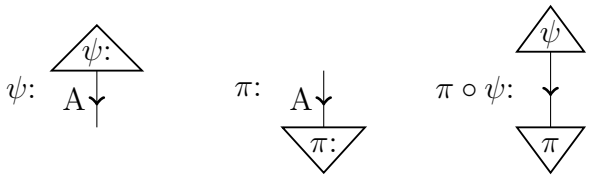


Note that in literature, the flow in the diagrams are sometimes upwards and sometimes downwards. Here I will use the convention that the information flows downwards. Inputs of the morphisms are at the top of the boxes and outputs are at the bottom.

Sequential composition is depicted just as one would expect, by connecting the output wire of the first morphism to the input wire of the second. The monoidal tensor is depicted as putting the two wires next to one another, and a morphism $g : A \otimes B \to C \otimes D$ as having two input and two output wires. The identity morphisms is depicted just as a wire. Here are some examples:

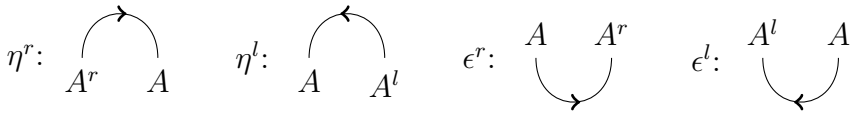


The unit object is depicted as empty space, so the states $\psi : I \to A$ are depicted as a box with no input wire and an output wire with type A. Effects are the dual of states, and they are of type $\pi : A \to I$ These take the graphical forms:

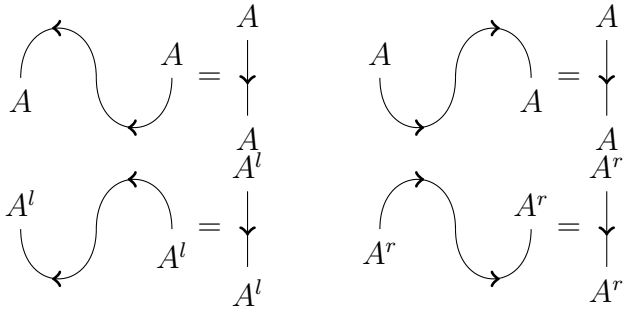


Graphical calculus for compact closed categories :

The maps $\eta^l, \eta^r, \epsilon^l$ and ϵ^r take the following forms in the graphical calculus:



The axioms of compact closure, referred to as the *snake identities* because of the visual form they take in the graphical calculus, are represented as follows:



More generally, the reduction rules for diagrammatic calculus allow continuous deformations. One such deformation that I will make use of is the *swing rule*:

Chapter 6

A Compositional Distributional Model

The model suggested in [4] uses the fact that both finite dimensional vector spaces and pregroups are compact closed categories, and builds a compositional distributional model by applying the maps obtained by pregroup reductions to vector spaces. The pregroup structure is used to specify the grammatical types of the words, and the reduction maps show how these grammatical types interact within a sentence. Each word also gets assigned a vector space as a distributional representation of its meaning. Since the maps obtained from the pregroup reduction are maps of a compact closed category, the corresponding maps for vector spaces are well defined. The application of these maps to the distributional representations of words offer a way to grammatically compose vector spaces, and to construct a distributional representation for the sentence. It has the additional advantage of ensuring that all the resulting sentence meanings live in the same space, making sentences with different grammatical structure comparable to each other.

6.1 Algebra of pregroups

Lambek [4] recently developed a new grammatical formalism called *pregroup grammars*, a type-categorial logic like his previous *Lambek calculus*.

Definition 6.1. [21] A **partially ordered monoid** $(P, \leq, \cdot, 1)$ consists of

- a set P,
- a monoid multiplication operator $\cdot : P \times P \to P$ satisfying the condition

$$(a \cdot b) \cdot c = a \cdot (b \cdot c) \text{ for all } a, b, c \in P$$

- and the monoidal unit $1 \in P$ where for all $a \in P$

$$a \cdot 1 = a = 1 \cdot a$$

- a partial order $\leq$ on P where

$$TODO$$

Definition 6.2. [21] A **pregroup** $(P, \leq, \cdot, 1, (-)^l, (-)^r)$ is a partially ordered monoid in which each element a has both a left adjoint a^l and a right adjoint a^r such that

$$a^l a \leq 1 \leq a a^l \text{ and } a a^r \leq 1 \leq a^r a$$

Adjoints of pregroups have the following properties [4] :

- Uniqueness: Adjoints are unique
- Order reversal: If $a \leq b$ then $b^r \leq a^r$ and $b^l \leq a^l$
- The unit is self adjoint: $1^l = 1 = 1^r$
- Multiplication operation is self adjoint: $(a \cdot b)^l = b^l \cdot a^l$ and $(a \cdot b)^r = b^r \cdot a^r$
- Opposite adjoints annihilate: $(a^r)^l = a = (a^l)^r$
- Same adjoints iterate: $a^{ll} a^l \leq 1 \leq a^{rr} a^r, a^{lll} a^{ll} \leq 1 \leq a^{rrr} a^{rr}, \ldots$

Both pregroup grammars and Lambek calculus are partially ordered monoids, but pregroup grammars replace the left and right adjoints for the monoid multiplication in Lambek calculus with left and right adjoints for elements.

If $a \leq b$, I will write $a \to b$ and say that a reduces to b. This terminology is useful when pregroups are applied to natural language, where each word gets assigned a pregroup type freely generated from a set of basic elements. The sentence is deemed to be grammatical if the concatenation of the types of the words reduce to the simple type of a sentence. As an example, consider a simple transitive sentence: "John likes Mary". "John" and "Mary" get assigned a basic type n for noun. "likes" is assigned a compound type, $(n^r s n^l)$. The role of a word with a compound type can be read of from the pregroup type assigned to it. In this example, "likes" takes a noun from the left and a noun from the right, and returns a sentence. The pregroup reduction for the sentence is:

$$n(n^r s n^l)n \to 1 s n^l n \to 1 s 1 \to s$$

Notice that the reduction sequence is not in general unique. It does not matter in this example whether we choose to reduce the left or the right side of the compound pregroup type first.

The basic types I will use are the following:

$n:$ noun	$s:$ declarative statement
$j:$ infinitive of the verb	$\sigma:$ glueing type

Pregroups as compact closed categories. A pregroup **P** is a concrete instance of a compact closed category. The underlying category is given by the partial order, and the monoidal structure is induced by the monoid structure of the pregroup. The $\eta^l, \eta^r, \epsilon^l, \epsilon^r$ maps are as follows:

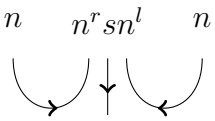
$$\eta^l = [1 \leq p \cdot p^l] \qquad \epsilon^l = [p^l \cdot p \leq 1]$$
$$\eta^r = [1 \leq p^r \cdot p] \qquad \epsilon^r = [p \cdot p^r \leq 1]$$

These satisfy the axioms of compact closure. Consider the first snake identity:

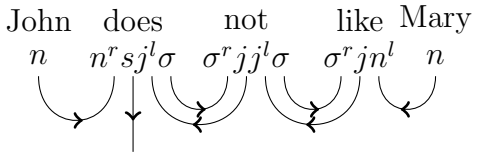
$$(1_p \otimes \epsilon^l_p) \circ (\eta^l_p \otimes 1_p):$$
$$p = 1p \leq pp^l p \leq p1 = p$$

The other snake identities are likewise easily confirmed.

Pregroup reductions in diagrammatic calculus. Since the pregroup reductions are maps of a compact closed category, they can be modeled by the graphical calculus. The graphical representation for a sentence of type "John likes Mary" is as following:



Here, one can see that the sentence is grammatical by observing that the only outgoing wire is of type s, the type of a sentence. All the other components cancel each other. The visualization is especially useful when the sentence structure is more complicated. For example, consider the sentence "John does not like Mary". The pregroup types for the words and the reduction is as follows:



6.2 Finite dimensional vector spaces

Finite dimensional vector spaces over the base field $\mathbb{R}$ together with linear maps, form a monoidal category, referred to as **FVect**. The monoidal tensor is the usual vector space tensor and the monoidal unit is the base field $\mathbb{R}$. It is also a compact closed category:

FVect as a compact closed category. The left and the right adjoint of an object V are both taken to be equal to V. The compact closure maps are defined as follows:

Given a vector space V with basis $\{\overrightarrow{e_i}\}_i$

$$\begin{aligned}\eta_V^l = \eta_V^r : \mathbb{R} &\to V \otimes V \\ 1 &\mapsto \sum_i e_i \otimes e_i \\ \epsilon_V^l = \epsilon_V^r : V \otimes V &\to \mathbb{R} \\ \sum_{ij} c_{ij}\ v_i \otimes w_i &\mapsto \sum_{ij} c_{ij} \langle v_i | w_i \rangle\end{aligned}$$

These satisfy the axioms of compact closure. Let $\overrightarrow{v} \in V$, and consider the first snake identity:

$$\begin{aligned}(1_V \otimes \epsilon_V^l) \circ (\eta_V^l \otimes 1_V) :: \overrightarrow{v} &\mapsto \left(\sum_i \overrightarrow{e_i} \otimes \overrightarrow{e_i} \right) \otimes \overrightarrow{v} \\ &\mapsto \sum_i \overrightarrow{e_i} \langle \overrightarrow{e_i} | \overrightarrow{v} \rangle = \overrightarrow{v}\end{aligned}$$

The other snake identities follow by a very similar calculation.

6.3 Categorical representation of meaning space

The tensor in **FVect** is commutative up to isomorphism. This causes the left and the right adjoints to be the same, and thus for the left and the right compact closure maps to coincide. This makes it impossible to express a map from the meanings of words to the meanings of sentences solely using the maps in **FVect**, since such a map is expected to return a different representation for the sentence "dog bit man" and "man bit dog". In other words, a symmetric compact closed category cannot take the effect of word ordering on meaning into account. [4] proposes a way around this obstacle by considering the product category $\mathbf{FVect} \times \mathbf{P}$ where $\mathbf{P}$ is the category given by a pregroup.

Objects in **FVect** are of the form (V, p), where V is the vector space representation of the meaning and p is the pregroup type. There exists a morphism $(f, \leq) : (V, p) \to (W, q)$ if there exists a morphism $f : V \to W$ in **FVect** and $p \leq q$ in **P**.

The compact closed structure of **FVect** and **P** lifts componentwise to the product category **FVect** $\times$ **P**:

$$\eta^l : (\mathbb{R}, 1) \to (V \otimes V, p \cdot p^l)$$

$$\eta^r : (\mathbb{R}, 1) \to (V \otimes V, p^r \cdot p)$$

$$\epsilon^l : (V \otimes V, p^l \cdot p) \to (\mathbb{R}, 1)$$

$$\epsilon^r : (V \otimes V, p \cdot p^r) \to (\mathbb{R}, 1)$$

Definition 6.3. An object (V, p) in the product category is called a **meaning space**, where V is the vector space in which the meanings $\overrightarrow{v} \in V$ of strings of type p live.

6.4 From the meanings of words to the meaning of the sentence map

The idea behind the from-meanings-of-words-to-the-meaning-of-the-sentence map is that the pregroup reductions guide the order in which the compact closure maps are applied to the vector spaces.

Definition 6.4. Let $v_1 v_2 \dots v_n$ be a string of words, each v_i with a meaning space representation $\overrightarrow{v_i} \in (V_i, p_i)$. Let $x \in P$ be a pregroup type such that $[p_1 p_2 \dots p_n \leq x]$ Then the meaning vector for the string is:

$$\overrightarrow{v_1 v_2 \dots v_n} \in (W, x) \equiv f(v_1 \otimes v_2 \otimes \dots \otimes v_n)$$

where f is defined to be the application of the compact closure maps obtained from the reduction $[p_1 p_2 \dots p_n \leq x]$ to the composite vector space $V_1 \otimes V_2 \otimes \dots \otimes V_n$.

This framework uses the maps of the pregroup reductions and the elements of objects in **FVect**. The diagrammatic calculus provides a tool to reason about both.

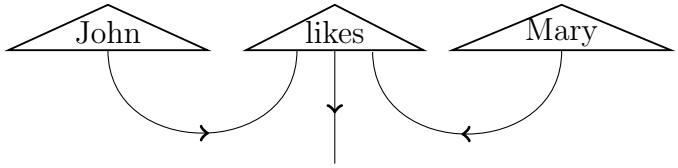
As an example, take the sentence "John likes Mary". It has the pregroup type $nn^r sn^l n$, and the vector representations $\overrightarrow{John}, \overrightarrow{Mary} \in V$ and $\overrightarrow{likes} \in V \otimes S \otimes V$. The morphism in **FVect** $\times$ **P** corresponding to the map is of type:

$$(V \otimes (V \otimes S \otimes V) \otimes V, nn^r sn^l n) \to (s, S)$$

From the pregroup reduction $[nn^r sn^l n \to s]$ we obtain the compact closure maps $\epsilon^r 1 \epsilon^l$. In **FVect** this translates into:

$$\epsilon_V \otimes 1_S \otimes \epsilon_V : V \otimes (V \otimes S \otimes V) \otimes V \to S$$

This map, when applied to $\overrightarrow{John} \otimes \overrightarrow{likes} \otimes \overrightarrow{Mary}$ has the following depiction in the diagrammatic calculus:



Note that this construction treats the verb "likes" essentially as a function that takes two inputs of type V, and outputs a vector of type S.

For the explicit calculation, consider how the vector space representation for "likes" look:

$$\overrightarrow{likes} = \sum_{ijk} c_{ijk} v_i \otimes s_j \otimes v_k$$

where v_i is an orthonormal basis for V and s_j is an orthonormal basis for S. Then

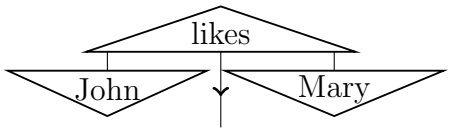
$$\overrightarrow{John\ likes\ Mary} = \epsilon_V \otimes 1_S \otimes \epsilon_V(\overrightarrow{John} \otimes \overrightarrow{likes} \otimes \overrightarrow{Mary}) \tag{6.1}$$

$$= \sum_{ijk} \langle John|v_i\rangle s_j \langle v_k|Mary\rangle \tag{6.2}$$

The reductions in diagrammatic calculus help reduce the final calculation to a simpler term. The non-reduced reduction, when expressed in dirac notation reads:

$$(\langle \epsilon_V^r| \otimes 1_S \otimes \langle \epsilon_V^l|) \circ |\overrightarrow{John} \otimes \overrightarrow{likes} \otimes \overrightarrow{Mary}\rangle$$

But we can "swing" $\overrightarrow{John}$ and $\overrightarrow{Mary}$ in accord with the reduction rules in the diagrammatic calculus. The diagram then reduces to:



This results in a simpler expression that needs to be calculated:

$$(\langle \overrightarrow{John}| \otimes 1_S \otimes \langle \overrightarrow{Mary}|) \circ |\overrightarrow{likes}\rangle$$

Chapter 7

Density Matrices as Elements of a Compact Closed Category

Recall that in **FVect**, vectors $|v\rangle \in V$ are in one-to-one correspondence with morphisms of type $v : I \to V$. Likewise, pure states of the form $|v\rangle\langle v|$ are in one-to-one correspondence with morphisms $v \circ v^\dagger : V \to V$ such that $v^\dagger \circ v = \mathrm{id}_I$ (notice that this corresponds to the condition that $\langle v|v\rangle = 1$). A general (mixed) state ρ is a positive morphism of the form $\rho : A \to A$. One can re-express the mixed states $\rho : A \to A$ as elements $\rho : I \to A^* \otimes A$. Here, as in the case for **FVect** as a compact closed category, I assign $A^* = A$.

Definition 7.1. f is a **completely positive map** if

1. It maps density matrices into density matrices. In other words, f is positive for any positive operator A.

2. Whenever it is tensored with the identity map on any space, it maps the density matrices in the combined space to density matrices: $(\ \mathrm{id}_V \otimes f)A$ is positive for any positive operator A and any space V.

Theorem 7.2. *completely positive maps in* **FVect** *form a monoidal category:*

- *The identity map* $id : A^* \otimes A \to A^* \otimes A$ *is completely positive.*

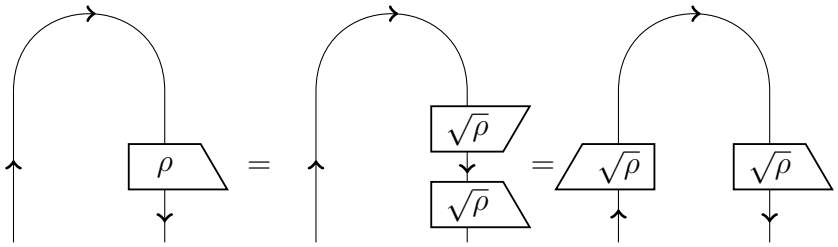
- *If* $f : A^* \otimes A \to B^* \otimes B$ *and* $g : B^* \otimes B \to C^* \otimes C$ *are completely positive maps, then* $g \circ f$ *is also completely positive.*

- *If* f *and* g *are completely positive maps, then* $f \otimes g$ *is also completely positive*

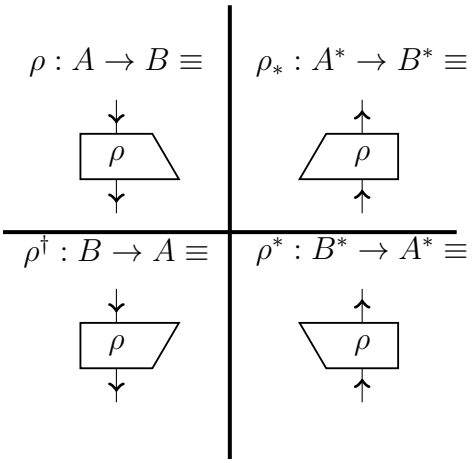
See [17] for the proof.

Thus one can define a new category **CPM**(**FVect**) where a morphism $A \to B$ in **CPM**(**FVect**) is a completely positive map $A^* \otimes A \to B^* \otimes B$ in **FVect**. The elements $I \to A$ in **CPM**(**FVect**) are of the form $I^* \otimes I \to B^* \otimes B$ in **FVect**, providing a monoidal category with density matrices as its elements.

CPM(FVect) in graphical calculus. A morphism $\rho : A \to A$ is positive if and only if there exists a map $\sqrt{\rho}$ such that $\rho = \sqrt{\rho}^\dagger \circ \sqrt{\rho}$. In **FVect**, the isomorphism between $\rho : A \to A$ and $\ulcorner \rho \urcorner : I \to A^* \otimes A$ is provided by $\eta^l = \eta^r$. The graphical representation of ρ in **FVect** then becomes:



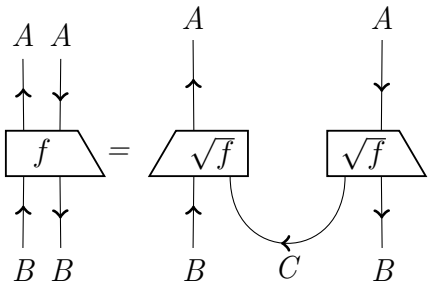
Here I use the convention that:



The graphical depiction of completely positive morphisms come from the following theorem:

Theorem 7.3. *(Stinespring Dilation Theorem) The following are equivalent:*

1. *$f : A^* \otimes A \to B^* \otimes B$ is completely positive*
2. *There is an object C and a morphism $\sqrt{f} : C \otimes B$ such that the following equation holds:*



$\sqrt{f}$ and C here are not unique. For the proof of the theorem see [17, p.149]

For a given density matrix ρ and a completely positive morphism f, one can easily check that the graphical representation of $f \circ \rho$ is indeed the graphical representation of a positive matrix.

CPM(FVect) as a compact closed category.

Theorem 7.4. CPM(FVect) *is a compact closed category where as in* **FVect**, $V^r = V^l = V$ *and the compact closure maps are defined to be:*

$$\eta^l = (\eta^r_{FVect} \otimes \eta^l_{FVect}) \circ (\mathbf{1}_A \otimes \sigma \otimes \mathbf{1}_A)$$

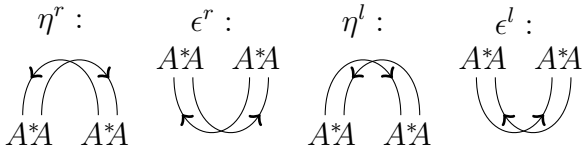
$$\eta^r = (\eta^l_{FVect} \otimes \eta^r_{FVect}) \circ (\mathbf{1}_A \otimes \sigma \otimes \mathbf{1}_A)$$

$$\epsilon^l = (\mathbf{1}_A \otimes \sigma \otimes \mathbf{1}_A) \circ (\epsilon^r_{FVect} \otimes \epsilon^l_{FVect})$$

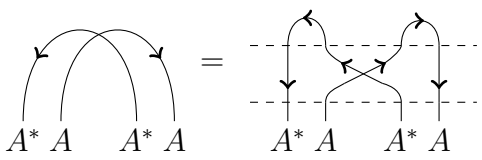
$$\epsilon^r = (\mathbf{1}_A \otimes \sigma \otimes \mathbf{1}_A) \circ (\epsilon^l_{FVect} \otimes \epsilon^r_{FVect})$$

where σ *is the swap map defined as* $\sigma(v \otimes w) = (w \otimes w)$.

Proof. The graphical construction of the compact closure maps boils down to doubling the objects and the wires:



Then adding some simple bends in the wire, we can see that these are in fact in the forms expressed above. Consider the diagram for η^r:



These clearly satisfy the axioms of compact closure since the components do. □

The concrete compact closure maps are as follows:

$$\eta^l = \eta^r : \mathbb{R} \to (V \otimes V) \otimes (V \otimes V) :: 1 \mapsto \sum_i \overrightarrow{e_i} \otimes \overrightarrow{e_i} \otimes \sum_j \overrightarrow{e_j} \otimes \overrightarrow{e_j}$$

$$\epsilon^l = \epsilon^r : (V \otimes V) \otimes (V \otimes V) \to \mathbb{R} :: \sum_{ijkl} c_{ijkl}\ \overrightarrow{v_i} \otimes \overrightarrow{w_j} \otimes \overrightarrow{u_k} \otimes \overrightarrow{p_l} \mapsto \sum_{ijkl} c_{ijkl}\ \langle \overrightarrow{v_i} | \overrightarrow{u_k} \rangle \langle \overrightarrow{w_j} | \overrightarrow{p_l} \rangle$$

Let ρ be a density operator defined on an arbitrary composite space $V_1 \otimes V_2 \otimes \ldots \otimes V_n$. then

$$\rho : V_1 \otimes V_2 \otimes \ldots \otimes V_n \to V_1 \otimes V_2 \otimes \ldots \otimes V_n$$

It has the density matrix representation:

$$\rho : I \to (V_1 \otimes V_2 \otimes \ldots \otimes V_n)^* \otimes (V_1 \otimes V_2 \otimes \ldots \otimes V_n)$$

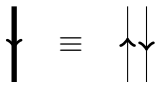
Since the underlying category FVect is symmetric, it has the swap map σ. This provides us with the isomorphism:

$$(V_1 \otimes V_2 \otimes \ldots \otimes V_n)^* \otimes (V_1 \otimes V_2 \otimes \ldots \otimes V_n) \sim (V_1^* \otimes V_1) \otimes (V_2^* \otimes V_2) \otimes \ldots \otimes (V_n^* \otimes V_n)$$

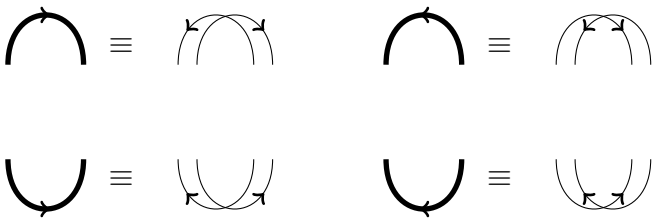
So ρ can be equivalently expressed as:

$$\rho : I \to (V_1^* \otimes V_1) \otimes (V_2^* \otimes V_2) \otimes \ldots \otimes (V_n^* \otimes V_n)$$

With this addition, we can simplify the diagrams used to express density matrices by using a single thick wire for the doubled wires:



Doubled compact closure maps can likewise be expressed by a single thick wire:



The diagrammatic expression of a from-meanings-of-words-to-the-meaning-of-the-sentence map using density matrices will therefore look exactly like the depiction of it in **FVect**, but with thick wires.

Chapter 8

Using Density Matrices to Model Meaning

Density matrices are probability distributions on subspaces of a Hilbert space, and one of their primary improvements on vector spaces is that they distinguish correlation and mixing. In vector space terminology, they distinguish between the case where we are certain that a state is an equally weighted sum of two basis elements, that this correlation is an intrinsic property of the state, and the case where we are simply lacking in knowledge whether the state is equal to one basis element or the other or any weighted sum of the two. The density matrix expressing the first case assigns probability 1 to the one dimensional vector space spanned by the line that is halfway through the two bases. If the two bases are x and y, this is the vector space spanned by $(\ 1\ \ 1\ )^T$. The density matrix expressing the second case is called a completely mixed state, since it assigns an equal probability to the entire two dimensional space. This is expressed by assigning equal probabilities to each basis vector. Assigning equal probabilities to any set of orthonormal basis vectors gives the completely mixed state.

If one wants to use the full power of density matrices in modeling meaning, one needs to establish an interpretation for the distinction between *mixing* and *correlation* in the context of linguistics. To make use of this distinction, it is necessary to take a step back from the co-occurrence space interpretation, where bases are defined to be chosen context words, and take a more abstract point of view. I will call the bases *contextual features*, and leave out their exact characterizations in this work as it is more of an implementation issue, and not an essential part of the mathematical framework I present here. I will just assume that these are the salient, quantifiable features of the contexts a word is observed in, and leave the characterization and extraction of these features for further work.

The simple co-occurrence model can be cast as a special case of this more general approach, where there is never any correlations between the basis vectors. Then all word meanings are mixtures of basis vectors, and they all commute with each other.

8.1 Characterizing similarity and representativeness

Density matrices allow us to define symmetric and asymmetric measures with relative naturality. In this section I define measures for similarity and representativeness on density matrices. Similarity is a symmetric measure for quantifying the closeness of two words in meaning. I survey three candidates for defining similarity over density matrices: *trace inner product*, *trace distance*, and *fidelity*. Trace inner product is the usual inner product defined on density matrices, while trace distance and fidelity are generalizations of distance measures on discrete probability distributions to density matrices. I conclude that fidelity has the most desirable properties between the three, but I include the other two since they might be more suitable for specific applications. I also suggest an asymmetric measure based on relative entropy to quantify representativeness between two density matrices. I will use representativeness to define conditions for inferring hyponymy-hypernymy relations.

Similarity. One of the fundamental aspects of a distributional semantic model is that it embodies an idea of meaning that is gradual, and comes equipped with a measure for the similarity. In vector space models the most widely used measure is cosine similarity [12], even though other measures are also used depending on the application [12, 5]

$$cosine(\overrightarrow{a}, \overrightarrow{b}) = \frac{\sum_i c_i^a c_i^b}{\sqrt{\sum_i (c_i^a)^2 \sum_i (c_i^b)^2}}$$

Where c_i^a and c_i^b are basis vectors for $\overrightarrow{a}$ and $\overrightarrow{b}$. When $\overrightarrow{a}$ and $\overrightarrow{b}$ are unit vectors, cosine similarity is equal to the inner product $\langle \overrightarrow{a} | \overrightarrow{b} \rangle$.

Trace inner product. Since density matrices also have an inner product, trace inner product between two density matrices is the first natural candidate for a similarity measure. Even though having an inner product on density matrices is very useful, One problem with using it as a measure of similarity is that it is

not always true that $\mathrm{tr}(A^T A)$ is equal to 1. Consider the completely mixed state $\rho = 1/2|0\rangle\langle 0| + 1/2|1\rangle\langle 1|$. Then $\mathrm{tr}(\rho^T \rho) = \mathrm{tr}(\rho^2) = 1/2$.

Another, perhaps bigger problem is that the trace distance between a state ρ and itself might be smaller than a trace distance between ρ and another state σ. To see this consider $\rho = 1/4|0\rangle\langle 0| + 3/4|1\rangle\langle 1|$, and $v = |1\rangle\langle 1|$. Then $\mathrm{tr}(\rho v) = 3/4 \geq \mathrm{tr}(\rho^2) = 5/8$. If trace inner product is taken to be the similarity measure, this would read that ρ is more similar to v then it is similar to itself.

Trace distance. This is one of the two most widely used distance measures in quantum mechanics:

Definition 8.1. The **trace distance** between two density matrices A and B is defined as

$$D(A, B) = \frac{1}{2}\mathrm{tr}|A - B|$$

where $|A| \equiv \sqrt{A^\dagger A}$.

Trace distance is a generalization of the *Kolmogorov distance* between two probability distributions $\{p_i\}$ and $\{q_i\}$:

$$K(p_i, q_i) \equiv \frac{1}{2}\sum_x |p_x - q_x|$$

When ρ and σ commute, the trace distance between σ and ρ is equal to the Kolmogorov distance between their eigenvalues.

Trace distance provides the condition that $D(\rho, \rho) = 0$, and the trace distance between two density matrices is never greater than 1. Since trace distance is a distance measure with values within the range $[0, 1]$, we can translate it into a similarity measure by defining

$$S'(\rho, \sigma) = 1 - D(\rho, \sigma)$$

Fidelity. This is the other widely used distance measure in quantum mechanics. It shares most of the desirable properties of trace distance, and has some additional advantages as a similarity measure for meaning.

Definition 8.2. The **fidelity** of two probability distributions $\{p_x\}$ and $\{q_x\}$ is defined by:

$$F(p_x, q_x) \equiv \sum_x \sqrt{p_x q_x}$$

The fidelity of two density operators ρ and σ is defined to be

$$F(\rho, \sigma) \equiv \mathrm{tr}\sqrt{\rho^{1/2}\sigma\rho^{1/2}}$$

Some useful properties of fidelity are:

1. It is a symmetric measure: $F(\rho, \sigma) = F(\sigma, \rho)$
2. $0 \leq F(\rho, \sigma) \leq 1$
3. $F(\rho, \sigma) = 1$ if and only if $\rho = \sigma$
4. If ρ and σ commute, then the fidelity of ρ and σ is equal to the classical fidelity of their eigenvalues.
5. If $|\psi\rangle\langle\psi|$ is a pure state and ρ is an arbitrary state,

$$F(|\psi\rangle\langle\psi|, \rho) = \sqrt{\langle\psi|\rho|\psi\rangle}$$

   That is, fidelity is equal to the square root of the overlap between ρ and the pure state.
6. If $|\phi\rangle\langle\phi|$ and $|\psi\rangle\langle\psi|$ are two pure states, their fidelity is equal to $|\langle\phi|\psi\rangle|$.

Fidelity and trace distance are related by the following formula [30]:

$$1 - F(\rho, \sigma) \leq D(\rho, \sigma) \leq \sqrt{1 - F(\rho, \sigma)^2}$$

So fidelity and trace distance are qualitatively equivalent, but using fidelity as a measure of semantic similarity has the advantages of guaranteeing that if two words are represented as projections onto one dimensional subspaces, their similarity value will be equal to the usual cosine similarity of the vectors.

Representativeness. The idea of using relative entropy to model hyponymy is as follows: assume that you are given as many sentences as you like where the word *dog* appears in, but the word dog is crossed out, and you are asked whether you think the crossed out word may be *animal* or not. Now because *animal* subsumes *dog*, it can be used in any context *dog* can be used in, thus you cannot ever be sure that the crossed out word is not *animal*. However, consider the other way around where the crossed out word is *animal* and you are asked whether you think the crossed out word is *dog* or not. Then it is likely that you will eventually come across a sentence of the sort: *"The — flapped its wings and flew away."* from which you immediately know that the crossed out word is not *dog*. Thus the distinguisability of one word from another given its usual contexts provide us a good metric for hyponymy.

A distributional hypothesis for hyponymy: The meaning of a word w subsumes the meaning of v if and only if it is appropriate to use w in all the contexts v is used.

This is a slightly more general version of the *Distributional Inclusion Hypothesis* stated in [19]:

> If u is semantically narrower than v, then a significant number of salient distributional features of u are also included in the feature vector of v.

The difference lies in the additional power the density matrix formalism provides: the distinction between mixing and correlation. The Distributional Inclusion Hypothesis considers only whether or not the target word occurs together with the salient distributional feature at all, and ignores any possible statistically significant correlations of features.

Note that [11] show that while there is ample evidence for the distributional inclusion hypothesis, this in itself does not necessarily provide a method to detect hyponymy-hypernymy pairs: the inclusion of a significant number of salient distributional features of u in v is not a reliable predictor that u implies v. One of their suggestions for improvement is to consider more than one word in the features, which is equivalent to taking correlations into account in a co-occurrence space where the bases are context words.

A measure for representativeness based on relative entropy. Relative entropy quantifies the distinguishability of one distribution from another. It is therefore a good candidate to base a measure of representativeness on.

Definition 8.3. the **representativeness** between ρ and σ is:

$$R(\rho, \sigma) = \frac{1}{1 + N(\rho||\sigma)}$$

Where $N(\rho||\sigma)$ is the quantum relative entropy between ρ and σ.

Notice that when the space is a co-occurrence space with no correlations of basis features, then ρ and σ commute, making quantum relative entropy of ρ and σ equal to the Kullback-Leibler divergence. Then the representativeness between ρ and σ is:

$$R(\rho, \sigma) = \frac{1}{1 + K(\rho||\sigma)}$$

Where $K(\rho||\sigma)$ is the Kullback-Leibler divergence.

Proposition 8.4. *For all density matrices ρ and σ, $R(\rho,\sigma) \leq 1$ with equality if and only if $\rho = \sigma$, and $0 \leq R(\rho,\sigma)$ with equality if and only if $supp(\rho) \cap ker(\sigma) \neq 0$*

Proof. The first part of the proposition follows directly from proposition 4.6, and the second part from proposition 4.7. □

The second property reflects the idea that if there is a context in which it is appropriate to use v but not w, v is perfectly distinguishable from w. Such contexts are exactly those that fall within $supp(\rho) \cap ker(\sigma)$.

Characterizing hyponyms The quantitative measure on density matrices given by representativeness provide a qualitative partial order on meaning representations as follows:

$$\rho \precsim \sigma \text{ if } R(\rho,\sigma) > 0$$

$$\rho \sim \sigma if \rho \precsim \sigma \text{ and } \sigma \precsim \rho$$

I will also denote a strict subsumption relation between ρ and σ by $\prec$:

$$\rho \prec \sigma \text{ if } R(\rho,\sigma) > 0 \text{ and } R(\sigma,\rho) = 0$$

Corollary 8.5. *The following are equivalent:*

1. $\rho \precsim \sigma$
2. $supp(\rho) \subseteq supp(\sigma)$
3. *There exists a positive operator ρ' and $p > 0$ such that $\sigma = p\rho + \rho'$*

Proof. (1) $\Rightarrow$ (2) and (2) $\Rightarrow$ (1) follow directly from proposition 4.6.

(2) $\Rightarrow$ (3) since $supp(\rho) \subseteq supp(\sigma)$ implies that there exists a $p > 0$ such that $\sigma - p\rho$ is positive. Setting $\rho' = \sigma - p\rho$ gives the desired equality.

(3) $\Rightarrow$ (2) since $p > 0$, and so $supp(\sigma) = supp(p\rho + \rho') \subseteq supp(\rho)$. □

The equivalence relation $\sim$ groups any two density matrices ρ and σ with $supp(\rho) = supp(\sigma)$ into the same equivalence class, thus maps the set of density matrices on a Hilbert space $\mathcal{H}$ onto the set of projections on $\mathcal{H}$. The projections are in one-to-one correspondence with the subspaces of $\mathcal{H}$ and they form an orthomodular lattice, providing a link to the logical structure of the Hilbert space [41] aims to exploit by using density matrices in IR.

Let $\ulcorner w \urcorner$ and $\ulcorner v \urcorner$ be density matrix representations of the words v and w. Then v is a hyponym of w in this model if $v \prec w$.

Notice that even though this ordering on density matrices extracts a *yes/no* answer for the question "is v a hyponym of w?" from the representativeness measure, the existence of the quantitative measure also lets us quantify the extent to which v is a hyponym of w. This provides with some flexibility in characterizing hyponymy through density matrices in practice. Instead of calling v a hyponym of w even when $R(\ulcorner v \urcorner, \ulcorner w \urcorner)$ gets arbitrarily small, one can require the representativeness to be above a certain threshold ϵ. This modification, however, has the down side of causing the transivity of hyponymy to fail.

8.2 From meanings of words to the meanings of sentences passage.

Recall that in chapter 6, a product category $\mathbf{P} \times \mathbf{FVect}$ is defined as the category of meaning spaces, and the from-the-meanings-of-words-to-the-meaning-of-the-sentence map is defined on this category as the application of the compact closure maps obtained from the pregroup reductions to vector spaces. As in the case for $\mathbf{FVect} \times \mathbf{P}$, $\mathbf{CPM}(\mathbf{FVect}) \times \mathbf{P}$ is a compact closed category, where the compact closure maps of $\mathbf{CPM}(\mathbf{FVect})$ and $\mathbf{P}$ lift component wise to the product category.

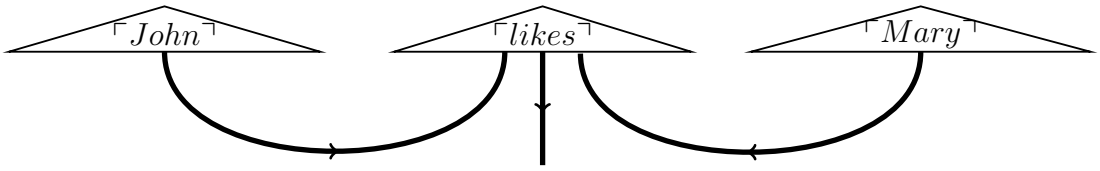
Definition 8.6. A **meaning space** in this new category is a pair $(V^* \otimes V, p)$ where $V^* \otimes V$ is the space in which density matrices $v : I \to V^* \otimes V$ of the pregroup type p live.

Definition 8.7. Let $v_1 v_2 \dots v_n$ be a string of words, each v_i with a meaning space representation $\ulcorner v_i \urcorner \in (V_i^* \otimes V, p_i)$. Let $x \in P$ be a pregroup type such that $[p_1 p_2 \dots p_n \leq x]$. Then the meaning density matrix for the string is defined as:

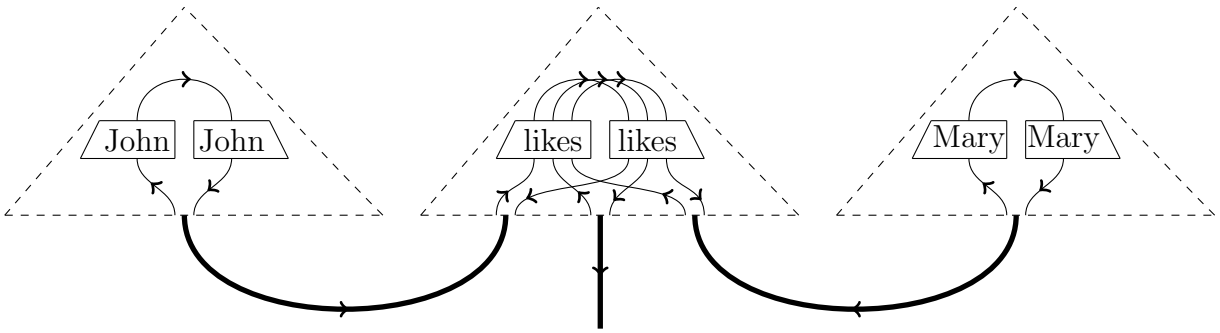
$$\ulcorner v_1 v_2 \dots v_n \urcorner \in (W^* \otimes W, x) \equiv f(v_1 \otimes v_2 \otimes \dots \otimes v_n)$$

where f is defined to be the application of the compact closure maps obtained from the reduction $[p_1 p_2 \dots p_n \leq x]$ to the composite density matrix space $(V_1 \otimes V_1^*) \otimes (V_2^* \otimes V_2) \otimes \dots \otimes (V_n^* \otimes V_n)$.

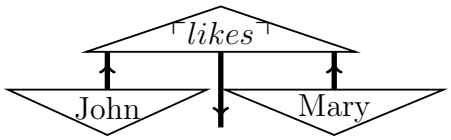
From a high level perspective, the reduction diagrams for $\mathbf{CPM}(\mathbf{FVect}) \times \mathbf{P}$ look no different than the original diagrams for $\mathbf{FVect} \times \mathbf{P}$, except that we depict it with thick instead of thin wires. Consider the previous example: "John likes Mary". As as in chapter 6 it has the pregroup type $n(n^r s n^l)n$, and the compact closure maps obtained from the pregroup reduction is $(\epsilon^r \otimes 1 \otimes \epsilon^l)$. The reduction diagram is:



However, we can also depict the internal anatomy of the density representations in **FVect**:



The graphical reductions for compact closed categories can be applied to the diagram:



Diagrammatically establishing the following equality:

$$(\epsilon^r \otimes 1 \otimes \epsilon^l)(\ulcorner John \urcorner \otimes \ulcorner likes \urcorner \otimes \ulcorner Mary \urcorner) = (\ulcorner John \urcorner \otimes 1 \otimes \ulcorner Mary \urcorner) \circ \ulcorner likes \urcorner$$

8.3 A hierarchy on sentences

One expects that if a word in a sentence is replaced by its hyponym, then the meanings of the original and the modified sentences would also have a relation akin to hyponymy. The following proposition shows that the sentence meaning map for simple transitive sentences achieves exactly that:

Proposition 8.8. *If $\rho, \sigma \in (N^* \otimes N, n)$, and $\rho \precsim \sigma$, then for density matrix $\ulcorner A \urcorner$ with a pregroup type n and $\ulcorner B \urcorner$ with a pregroup type $n^l s n^r$,*

$$f(\rho \otimes \ulcorner B \urcorner \otimes \ulcorner A \urcorner) \precsim f(\sigma \otimes \ulcorner B \urcorner \otimes \ulcorner A \urcorner) \text{ and}$$

$$f(\ulcorner A \urcorner \otimes \ulcorner B \urcorner \otimes \rho) \precsim f(\ulcorner A \urcorner \otimes \ulcorner B \urcorner \otimes \sigma)$$

where f is the from-meanings-of-words-to-the-meaning-of-the-sentence map in definition 8.7

Proof. If $\rho \precsim \sigma$, then there exists a positive operator ρ' and $\alpha > 0$ such that $\sigma = \alpha\rho + \rho'$ by proposition 8.5. Then

$$\begin{aligned} f(\sigma \otimes \ulcorner B \urcorner \otimes \ulcorner A \urcorner) &= (\epsilon^r \otimes 1 \otimes \epsilon^l)(\sigma \otimes \ulcorner B \urcorner \otimes \ulcorner A \urcorner) \\ &= (\sigma \otimes 1 \otimes \ulcorner B \urcorner) \circ \ulcorner A \urcorner \\ &= ((\alpha\rho + \rho') \otimes 1 \otimes \ulcorner B \urcorner) \circ \ulcorner A \urcorner \\ &= (\alpha\rho \otimes 1 \otimes \ulcorner B \urcorner) \circ \ulcorner A \urcorner + (\rho' \otimes 1 \otimes \ulcorner B \urcorner) \circ \ulcorner A \urcorner \\ f(\rho \otimes \ulcorner B \urcorner \otimes \ulcorner A \urcorner) &= (\rho \otimes 1 \otimes \ulcorner B \urcorner) \circ \ulcorner A \urcorner \end{aligned}$$

since $\alpha \neq 0$, $supp(f(\rho \otimes \ulcorner B \urcorner \otimes \ulcorner A \urcorner)) \subseteq supp(f(\sigma \otimes \ulcorner B \urcorner \otimes \ulcorner A \urcorner)$, which by proposition 8.5 proves the first part of the proposition. The second part is proved by a very similar calculation. □

In some cases, the more general sentence entails the more specific one: "Humans use language" entails "Polynesians use language", and "Jane loves herbs" entails "Jane loves basil". However, it is also common that the exact opposite is true: "Jane wants a beer" entails "Jane wants a beverage", and "A dog barked" entails "An animal barked". These examples illustrate that even between the simplest of sentences, the entailment relation relies on implicit quantifiers. In the next chapter I present some toy some examples that suggest some notion of entailment, but leave its rigorous characterization for further work.

Chapter 9

Examples

Here I will present several examples that demonstrate the application of the from-meanings-of-words-to-the-meaning-if-the-sentence map where the initial meaning representations of words are density matrices, and explore how the hierarchy on nouns induced by their density matrix representations carry over to a hierarchy in the sentence space.

Let "lions", "sloths". "plants" and "meat" have one dimensional representations in the noun space of our model:

$$\ulcorner lions \urcorner = |\overrightarrow{lions}\rangle\langle\overrightarrow{lions}|$$

$$\ulcorner sloths \urcorner = |\overrightarrow{sloths}\rangle\langle\overrightarrow{sloths}|$$

$$\ulcorner meat \urcorner = |\overrightarrow{meat}\rangle\langle\overrightarrow{meat}|$$

$$\ulcorner plants \urcorner = |\overrightarrow{plants}\rangle\langle\overrightarrow{plants}|$$

Let the representation of "mammals" be a mixture of one dimensional representations of individual animals:

$$\ulcorner mammals \urcorner = 1/2|\overrightarrow{lions}\rangle\langle\overrightarrow{lions}| + 1/2|\overrightarrow{sloths}\rangle\langle\overrightarrow{sloths}|$$

Notice that

$$\begin{aligned} N(\ulcorner lions \urcorner || \ulcorner mammals \urcorner) & \\ &= \mathrm{tr}(\ulcorner lions \urcorner \log \ulcorner lions \urcorner) - tr(\ulcorner lions \urcorner \log \ulcorner mammals \urcorner) \\ &= \log 1 - \frac{1}{2}\log\frac{1}{2} \\ &= 1 \end{aligned}$$

Hence $R(\ulcorner lions \urcorner, \ulcorner mammals \urcorner) = 1/2$.

$$
\begin{aligned}
&N(\ulcorner mammals \urcorner || \ulcorner lions \urcorner) \\
&\qquad = \mathrm{tr}(\ulcorner mammals \urcorner \log \ulcorner mammals \urcorner) - \mathrm{tr}(\ulcorner mammals \urcorner \log \ulcorner lions \urcorner) \\
&\qquad = \infty
\end{aligned}
$$

Since the intersection of the support of $\ulcorner mammals \urcorner$ and the kernel of $\ulcorner lions \urcorner$ is non-empty, so $R(\ulcorner mammals \urcorner, \ulcorner lions \urcorner) = 0$. This confirms that $\ulcorner lions \urcorner \prec \ulcorner mammals \urcorner$.

9.1 One dimensional truth theoretic sentences

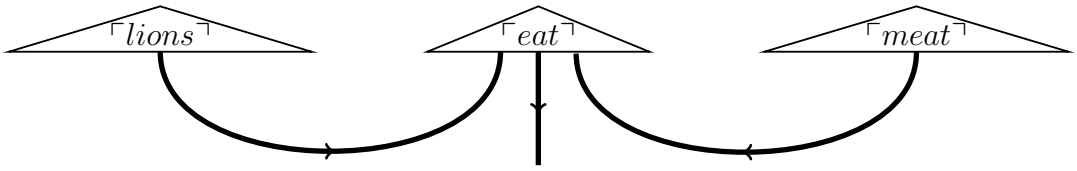
Consider a sentence space that is one dimensional, where 1 stands for true and 0 for false. Let

$$
\begin{aligned}
\ulcorner eat \urcorner =& (|\overrightarrow{sloths}\rangle|\overrightarrow{plants}\rangle + |\overrightarrow{lions}\rangle|\overrightarrow{meat}\rangle)(\langle\overrightarrow{sloths}|\langle\overrightarrow{plants}| + \langle\overrightarrow{lions}|\langle\overrightarrow{meat}|) \\
=& (|\overrightarrow{sloths}\rangle|\overrightarrow{plants}\rangle)(\langle\overrightarrow{sloths}|\langle\overrightarrow{plants}|) + \\
& (|\overrightarrow{sloths}\rangle|\overrightarrow{plants}\rangle)(\langle\overrightarrow{lions}|\langle\overrightarrow{meat}|) + \\
& (|\overrightarrow{lions}\rangle|\overrightarrow{meat}\rangle)(\langle\overrightarrow{sloths}|\langle\overrightarrow{plants}|) + \\
& (|\overrightarrow{lions}\rangle|\overrightarrow{meat}\rangle)(\langle\overrightarrow{lions}|\langle\overrightarrow{meat}|) \\
\sim& (|\overrightarrow{sloths}\rangle\langle\overrightarrow{sloths}| \otimes |\overrightarrow{plants}\rangle\langle\overrightarrow{plants}|) + \\
& (|\overrightarrow{sloths}\rangle\langle\overrightarrow{lions}| \otimes |\overrightarrow{plants}\rangle\langle\overrightarrow{meat}|) + \\
& (|\overrightarrow{lions}\rangle\langle\overrightarrow{sloths}| \otimes |\overrightarrow{meat}\rangle\langle\overrightarrow{plants}|) + \\
& (|\overrightarrow{lions}\rangle\langle\overrightarrow{lions}| \otimes |\overrightarrow{meat}\rangle\langle\overrightarrow{meat}|)
\end{aligned}
$$

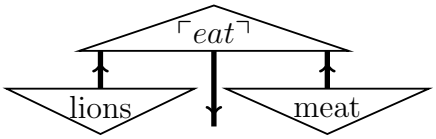
This is the density matrix representation of a pure composite state that relate "sloths" to "plants" and "lions" to "meat". If we fix the bases $\{\overrightarrow{lions}, \overrightarrow{sloths}\}$ for N_1, and $\{\overrightarrow{meat}, \overrightarrow{plants}\}$ for N_2, $\ulcorner eat \urcorner : N_1 \otimes N_1 \to N_2 \otimes N_2$ has the following matrix representation:

$$
\begin{pmatrix}
1 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1
\end{pmatrix}
$$

"Lions eat meat" . This is a transitive sentence, so as before, it gets assigned the pregroup type: $nn^l sn^r n$. The diagrammatic expression of the pregroup reduction is as follows:



This reduces to:



Explicit calculation gives:

$$
\begin{aligned}
&(\epsilon^l_N \otimes 1_S \otimes \epsilon^r_N)(\ulcorner lions \urcorner \otimes \ulcorner eat \urcorner \otimes \ulcorner meat \urcorner) \\
&\qquad = \langle \overrightarrow{lions} | \overrightarrow{sloths} \rangle^2 \langle \overrightarrow{plants} | \overrightarrow{meat} \rangle^2 + \\
&\qquad\quad \langle \overrightarrow{lions} | \overrightarrow{sloths} \rangle \langle \overrightarrow{lions} | \overrightarrow{lions} \rangle \langle \overrightarrow{meat} | \overrightarrow{meat} \rangle \langle \overrightarrow{plants} | \overrightarrow{meat} \rangle + \\
&\qquad\quad \langle \overrightarrow{lions} | \overrightarrow{lions} \rangle \langle \overrightarrow{lions} | \overrightarrow{sloths} \rangle \langle \overrightarrow{meat} | \overrightarrow{meat} \rangle \langle \overrightarrow{plants} | \overrightarrow{meat} \rangle + \\
&\qquad\quad \langle \overrightarrow{lions} | \overrightarrow{lions} \rangle^2 \langle \overrightarrow{meat} | \overrightarrow{meat} \rangle^2 \\
&\qquad = 0 + 0 + 0 + 1 \\
&\qquad = 1
\end{aligned}
$$

"Sloths eat meat" . This sentence has a very similar calculation to the one above, and has the result:

$$(\epsilon^l_N \otimes 1_S \otimes \epsilon^r_N)(\ulcorner sloths \urcorner \otimes \ulcorner eat \urcorner \otimes \ulcorner meat \urcorner) = 0$$

"Mammals eat meat" . This sentence has the same pregroup types as the first sentence, and so has the same reduction map:

$$
\begin{aligned}
&(\epsilon^l_N \otimes 1_S \otimes \epsilon^r_N)(\ulcorner mammals \urcorner \otimes \ulcorner eat \urcorner \otimes \ulcorner meat \urcorner) \\
&\qquad = (\epsilon^l_N \otimes 1_S \otimes \epsilon^r_N)((\frac{1}{2}\ulcorner lions \urcorner + \frac{1}{2}\ulcorner sloths \urcorner) \otimes \ulcorner eat \urcorner \otimes \ulcorner meat \urcorner) \\
&\qquad = \frac{1}{2}(\epsilon^l_N \otimes 1_S \otimes \epsilon^r_N)(\ulcorner lions \urcorner \otimes \ulcorner eat \urcorner \otimes \ulcorner meat \urcorner) + \\
&\qquad\quad \frac{1}{2}(\epsilon^l_N \otimes 1_S \otimes \epsilon^r_N)(\ulcorner sloths \urcorner \otimes \ulcorner eat \urcorner \otimes \ulcorner meat \urcorner) \\
&\qquad = \frac{1}{2}
\end{aligned}
$$

The meaning for the sentence "Mammals eat meat" is a mixture of "lions eat meat", which is true, and "sloths eat meat" which is false. Thus the value 1/2 can be interpreted as being neither completely true or completely false: the sentence "mammals eat meat" is true for certain mammals and false for others.

9.2 Two dimensional truth theoretic sentence

For two dimensional truth theoretic meaning, I will keep the noun representations the same, but use a two dimensional sentence space where:

$$true \equiv |0\rangle \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ and } false \equiv |1\rangle \equiv \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

This changes the representation of "eats". Let $A = \{lions, sloths\}$ and $B = \{meat, plants\}$

$$\ulcorner eat \urcorner \equiv \sum_{\substack{a_1,a_2 \in A \\ b_1,b_2 \in B}} |\overrightarrow{a_1}\rangle\langle\overrightarrow{a_2}| \otimes |\overrightarrow{x}\rangle\langle\overrightarrow{x}| \otimes |\overrightarrow{b_1}\rangle\langle\overrightarrow{b_2}|$$

where

$$|x\rangle \equiv \begin{cases} |0\rangle \text{ if } |a_1\rangle|b_1\rangle, |a_2\rangle|b_2\rangle \in \{|\overrightarrow{lions}\rangle|\overrightarrow{meat}\rangle, |\overrightarrow{sloths}\rangle|\overrightarrow{plants}\rangle\} \\ |1\rangle \text{ otherwise} \end{cases}$$

The generalized matrix representation of this verb in the spirit of [**?**] is:

$$\left(\begin{array}{cccc|cccc} 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \end{array}\right)$$

"Lions eat meat" The calculation of the meaning of the sentence is almost exactly the same as the case for one dimensional meaning, only the result is not the scalar that stands for *true* but the density matrix:

$$(\epsilon_N^l \otimes 1_S \otimes \epsilon_N^r)(\ulcorner lions \urcorner \otimes \ulcorner eat \urcorner \otimes \ulcorner meat \urcorner) = |0\rangle\langle 0|$$

"Sloths eat meat" Likewise, the calculation for "Sloths eat meat" return $false$:

$$(\epsilon_N^l \otimes 1_S \otimes \epsilon_N^r)(\ulcorner sloths \urcorner \otimes \ulcorner eat \urcorner \otimes \ulcorner meat \urcorner) = |1\rangle\langle 1|$$

"Mammals eat meat" As we have seen before, "Mammals eat meat" has the meaning that is the mixture of "Lions eat meat" and "Sloths eat meat":

$$\begin{aligned}(\epsilon_N^l \otimes 1_S \otimes \epsilon_N^r)(\ulcorner mammals \urcorner \otimes \ulcorner eat \urcorner \otimes \ulcorner meat \urcorner) &= \\ &= \frac{1}{2}(\epsilon_N^l \otimes 1_S \otimes \epsilon_N^r)(\ulcorner lions \urcorner \otimes \ulcorner eat \urcorner \otimes \ulcorner meat \urcorner) + \\ &\quad \frac{1}{2}(\epsilon_N^l \otimes 1_S \otimes \epsilon_N^r)(\ulcorner sloths \urcorner \otimes \ulcorner eat \urcorner \otimes \ulcorner meat \urcorner) \\ &= \frac{1}{2}|1\rangle\langle 1| + \frac{1}{2}|0\rangle\langle 0|\end{aligned}$$

So in a two dimensional truth theoretic model, "Mammals eat meat" give the completely mixed state in the sentence space, which has maximal entropy. This is equivalent to saying that we have no real knowledge whether mammals in general eat meat or not. Even if we are completely certain about whether individual mammals that span our space for "mammals" eat meat, this information differs uniformly within the members of the class, so we cannot generalize.

Already with a two dimensional truth theoretic model, the relation $\ulcorner lions \urcorner \prec \ulcorner mammals \urcorner$ carries over to the sentences:

$$\begin{aligned}N(\ulcorner \text{lions eat meat} \urcorner || \ulcorner \text{mammals eat meat} \urcorner) & \\ &= N\left(|0\rangle\langle 0| \;\middle|\middle|\; \frac{1}{2}|0\rangle\langle 0| + \frac{1}{2}|1\rangle\langle 1|\right) \\ &= (|0\rangle\langle 0|)\log(|0\rangle\langle 0|) - (|0\rangle\langle 0|)\log\left(\frac{1}{2}|0\rangle\langle 0| + \frac{1}{2}|1\rangle\langle 1|\right) \\ &= 1\end{aligned}$$

$N(\ulcorner \text{mammals eat meat} \urcorner || \ulcorner \text{lions eat meat} \urcorner) = \infty$ since the intersection of the support of the first argument and the kernel of the second argument is non-trivial. Thus the representativeness of the sentences are as follows:

$$R(\ulcorner \text{lions eat meat} \urcorner, \ulcorner \text{mammals eat meat} \urcorner) = 1/2$$

$$R(\ulcorner \text{mammals eat meat} \urcorner || \ulcorner \text{lions eat meat} \urcorner) = 0$$

$$\ulcorner \text{lions eat meat} \urcorner \prec \ulcorner \text{mammals eat meat} \urcorner$$

The from-meaning-of-words-to-the-meaning-of-the-sentence map carries the hyponymy relation in the subject words of the respective sentences to the resulting sentence meanings. By using the density matrix representations of word meanings

together with the categorical map from the meanings of words to the meanings of sentences, the knowledge that a lion is an animal lets us infer that "mammals eat meat" implies "lions eat meat":

$$(\ulcorner lions \urcorner \prec \ulcorner mammals \urcorner) \rightarrow (\ulcorner \text{lions eat meat} \urcorner \prec \ulcorner \text{mammals eat meat} \urcorner)$$

"Dogs eat meat" To see how the completely mixed state differs from a perfectly correlated but pure state in the context of linguistic meaning, consider a new noun $\ulcorner dog \urcorner = |\overrightarrow{dog}\rangle\langle\overrightarrow{dog}|$ and redefine eat in terms of the bases $\{\overrightarrow{lions}, \overrightarrow{dogs}\}$ and $\{\overrightarrow{meat}, \overrightarrow{plants}\}$, so that it will reflect the fact that dogs eat *both* meat and plants. I define "eat" so that it results in the value of being "half-true half-false" when it takes "dogs" as subject and "meat" or "plants" as object. The value "half-true half-false" is the superposition of $true$ and $false$: $\frac{1}{2}|0\rangle + \frac{1}{2}|1\rangle$

$\ulcorner eat \urcorner$ will still be a pure state so that we can first write the representation of "eat" in **FVect**:

$$\begin{aligned}|\overrightarrow{eat}\rangle =& |\overrightarrow{lions}\rangle \otimes |0\rangle \otimes |\overrightarrow{meat}\rangle + \\ & |\overrightarrow{lions}\rangle \otimes |1\rangle \otimes |\overrightarrow{plants}\rangle + \\ & |\overrightarrow{dogs}\rangle \otimes (\frac{1}{2}|0\rangle + \frac{1}{2}|1\rangle) \otimes |\overrightarrow{meat}\rangle + \\ & |\overrightarrow{dogs}\rangle \otimes (\frac{1}{2}|0\rangle + \frac{1}{2}|1\rangle) \otimes |\overrightarrow{plants}\rangle\end{aligned}$$

The density matrix representation of "eat" then becomes:

$$\ulcorner eat \urcorner = |\overrightarrow{eat}\rangle\langle\overrightarrow{eat}|$$

The calculation is as follows:

$$\begin{aligned}&(\epsilon^l_N \otimes 1_S \otimes \epsilon^r_N)(\ulcorner dogs \urcorner \otimes \ulcorner eat \urcorner \otimes \ulcorner meat \urcorner) \\ &\quad = (\epsilon^l_N \otimes 1_S \otimes \epsilon^r_N)(|\overrightarrow{dogs}\rangle\langle\overrightarrow{dogs}| \otimes |\overrightarrow{eat}\rangle\langle\overrightarrow{eat}| \otimes |\overrightarrow{meat}\rangle\langle\overrightarrow{meat}|) \\ &\quad = (\frac{1}{2}|0\rangle + \frac{1}{2}|1\rangle)(\frac{1}{2}\langle 0| + \frac{1}{2}\langle 1|)\end{aligned}$$

So in this case, we are certain that it is half-true and half-false that dogs eat meat. This is in contrast with the completely mixed state we got from "Mammals eat meat", for which the truth or falsity of the sentence was entirely unknown.

"Mammals eat meat", again . Let "mammals" now be defined as:

$$\ulcorner mammals \urcorner = \frac{1}{2}\ulcorner lions \urcorner + \frac{1}{2}\ulcorner dogs \urcorner$$

The calculation for this sentence with the new definition of "mammals" and "eat" gives:

$$\begin{aligned}
&(\epsilon_N^l \otimes 1_S \otimes \epsilon_N^r)(\ulcorner mammals \urcorner \otimes \ulcorner eat \urcorner \otimes \ulcorner meat \urcorner) \\
&\qquad = \frac{1}{2}(\epsilon_N^l \otimes 1_S \otimes \epsilon_N^r)(\ulcorner lions \urcorner \otimes \ulcorner eat \urcorner \otimes \ulcorner meat \urcorner) + \\
&\qquad\quad \frac{1}{2}(\epsilon_N^l \otimes 1_S \otimes \epsilon_N^r)(\ulcorner dogs \urcorner \otimes \ulcorner eat \urcorner \otimes \ulcorner meat \urcorner) \\
&\qquad = \frac{1}{2}|0\rangle\langle 0| + \frac{1}{2}((\frac{1}{2}|0\rangle + \frac{1}{2}|1\rangle)(\frac{1}{2}\langle 0| + \frac{1}{2}\langle 1|)) \\
&\qquad = \frac{3}{4}|0\rangle\langle 0| + \frac{1}{4}|0\rangle\langle 1| + \frac{1}{4}|1\rangle\langle 0| + \frac{1}{4}|1\rangle\langle 1|
\end{aligned}$$

This time the resulting sentence representation is not completely mixed. This means that we can generalize the knowledge we have from the specific instances of mammals to the entire class to some extent, but still we cannot generalize completely. This is a mixed state, which indicates that even if the sentence is closer to $true$ than to $false$, the degree of truth isn't homogeneous throughout the elements of the class. The non-zero non-diagonals indicate that it is also partially correlated, which means that there are some instances of "mammals" for which this sentence is true to a degree, but not completely. The relative similarity measures of $true$ and $false$ to the sentence can be calculated explicitly using fidelity:

$$F\big(|1\rangle\langle 1|, \ulcorner \text{mammals eat meat} \urcorner\big) = \langle 1|\ulcorner \text{mammals eat meat} \urcorner|1\rangle = \frac{1}{4}$$

$$F\big(|0\rangle\langle 0|, \ulcorner \text{mammals eat meat} \urcorner\big) = \langle 0|\ulcorner \text{mammals eat meat} \urcorner|0\rangle = \frac{3}{4}$$

Notice that these values are different than the values for the representativeness for $true$ and $false$ of the sentence, even thought they are proportional: the more representative a density matrix of another, the more similar they are to each other as well.

$$\begin{aligned}
&N\big(|1\rangle\langle 1| \parallel \ulcorner \text{mammals eat meat} \urcorner\big) \\
&\quad = \operatorname{tr}\big(|1\rangle\langle 1| \log(|1\rangle\langle 1|)\big) - \operatorname{tr}\big(|1\rangle\langle 1| \log(\frac{3}{4}|0\rangle\langle 0| + \frac{1}{4}|0\rangle\langle 1| + \frac{1}{4}|1\rangle\langle 0| + \frac{1}{4}|1\rangle\langle 1|)\big) \\
&\quad \approx 2
\end{aligned}$$

$R\big(|1\rangle\langle 1| \parallel \ulcorner\text{mammals eat meat}\urcorner\big) \approx .33$

$$\begin{aligned} &N\big(|0\rangle\langle 0| \parallel \ulcorner\text{mammals eat meat}\urcorner\big) \\ &\quad = \mathrm{tr}\big(|0\rangle\langle 0|)\log(|0\rangle\langle 0|)\big) - \mathrm{tr}\big(|0\rangle\langle 0|\log(\frac{3}{4}|0\rangle\langle 0| + \frac{1}{4}|0\rangle\langle 1| + \frac{1}{4}|1\rangle\langle 0| + \frac{1}{4}|1\rangle\langle 1|)\big) \\ &\quad \approx 0.41 \end{aligned}$$

$R\big(|0\rangle\langle 0| \parallel \ulcorner\text{mammals eat meat}\urcorner\big) \approx 0.71$

9.3 Verb hierarchy

Just as a meaning of a noun can subsume the meaning of another noun, a meaning of a verb might subsume the meaning of another verb. This is called *troponymy* and it is one of the relations in WordNet. An example of troponymy is the relation between "slap" and "hit".

"Mary hit Jane" . Let the noun space be the same for both the subject and the object, and be spanned by basis vectors $\{\overrightarrow{Mary}, \overrightarrow{Jane}\}$. Let it be true that Mary slapped Jane, but she didn't punch or kick her. Jane, on the other hand, both kicked and punched Mary back. Then, in a one dimensional truth theoretic model the vector space representations of these verbs are:

$$|\overrightarrow{slap}\rangle \equiv |\overrightarrow{Mary}\rangle|\overrightarrow{Jane}\rangle$$

$$|\overrightarrow{punch}\rangle \equiv |\overrightarrow{kick}\rangle \equiv |\overrightarrow{Jane}\rangle|\overrightarrow{Mary}\rangle$$

The density matrix representations are:

$$\ulcorner punch\urcorner = |\overrightarrow{punch}\rangle\langle\overrightarrow{punch}| \qquad \ulcorner kick\urcorner = |\overrightarrow{kick}\rangle\langle\overrightarrow{kick}| \qquad \ulcorner slap\urcorner = |\overrightarrow{slap}\rangle\langle\overrightarrow{slap}|$$

and

$$\ulcorner hit\urcorner = \ulcorner punch\urcorner + \ulcorner kick\urcorner + \ulcorner slap\urcorner$$

Then the calculation for "Mary hit Jane" is as follows:

$$\begin{aligned} &(\epsilon_N^l \otimes 1_S \otimes \epsilon_N^r)(\ulcorner Mary\urcorner \otimes \ulcorner hit\urcorner \otimes \ulcorner Jane\urcorner) \\ &\qquad = (\epsilon_N^l \otimes 1_S \otimes \epsilon_N^r)(\ulcorner Mary\urcorner \otimes \ulcorner slapped\urcorner \otimes \ulcorner Jane\urcorner) + \\ &\qquad\quad (\epsilon_N^l \otimes 1_S \otimes \epsilon_N^r)(\ulcorner Mary\urcorner \otimes \ulcorner punched\urcorner \otimes \ulcorner Jane\urcorner) + \\ &\qquad\quad (\epsilon_N^l \otimes 1_S \otimes \epsilon_N^r)(\ulcorner Mary\urcorner \otimes \ulcorner kicked\urcorner \otimes \ulcorner Jane\urcorner) \\ &\qquad = 1 + 0 + 0 = 1 \end{aligned}$$

Thus in a one dimensional truth theoretic setting, by defining the general verb through more specific ones we can get the truth value of a sentence where the general verb is used. Unfortunately, this construction fails in the two dimensional truth theoretic case with the truth interpretation, since the mixing causes the meaning of "Mary hit Jane" to be only one third true.

Chapter 10

Conclusion and Further Work

10.1 Summary

I show how density matrices can be used to model hyponymy relations if the following hypothesis is assumed:

Distributional hypothesis for hyponymy: The meaning of a word w subsumes the meaning of v if and only if it is appropriate to use w in all the contexts v is used.

I suggest to use density matrices to model hyponymy, but instead of using context words as bases, I assume a model that takes some abstract salient features of the context as a basis. This point of view characterizes a one dimensional subspace as a particular context, which stands for a particular correlation of feature elements. A general density matrix represents a probability distribution on correlations of feature elements.

Fidelity of two density matrices provide a similarity measure that coincides with the cosine similarity on two pure states. There are a number of measures based on Kullback-Leibler divergence in literature [22], and I suggest a measure of representativeness using quantum relative entropy that generalizes KL-divergence to density matrices. Relative entropy is used as a measure of distinguishability, and fits well with the distributional hypothesis for hyponymy.

The model where basis elements are chosen to be context words can be realized as a special case for the more general salient-features framework, where there is never any correlations between basis vectors. All density matrices are then mixtures of basis vectors, and they all commute with each other. In this case, the quantum relative entropy reduces to the classical KL-divergence.

The category of density matrices are elements in a compact closed category, and hence the from-meanings-of-words-to-the-meaning-of-the-sentence map can be used on density matrix representations, providing a composition method of word meanings into the meaning of the sentence that takes the grammar of the sentence into account. For a simple transitive sentence, this map is monotone in relation to the partial order on the density matrices obtained from the representativeness measure.

10.2 Discussion and Further Work

The model presented in this dissertation relies on the distributional hypothesis for hyponymy in a very strong sense: if it is appropriate for w to be used in all the contexts v is used, then the distributional model is assumed to reflect this perfectly. If $\ulcorner w \urcorner$ has even the slightest non-zero weight on a subspace that belongs to the kernel of $\ulcorner v \urcorner$, w will be judged to be perfectly distinguishable from v. Assuming such a perfect representation allows for a clearer theoretical framework, but it is impossible to obtain with statistical analysis of corpus data, thus unfit for implementation in its current form. [22] present two measures that have KL-divergence as their basis, and their generalizations to density matrices might overcome this issue.

There is also the question of how density matrix representations of meanings that make use of both mixing and correlation might be generated from corpus data. Here I sketch a skeleton for a possible algorithm:

A method to generate density matrix representations from corpus.

- Each particular context has a pure state representation, since it is a particular correlation of the feature elements.
- If a word w occurs in contexts that are close to each other (in terms of their pure state representations) to some statistically significant degree, these contexts are clustered together to a single average context representation.
- The average vector for each cluster gets assigned a weight according to the clusters relative frequency.
- These weights are generalized to a probability distribution on the entire space, providing a density matrix representation for contextual distribution of w.

This can be seen as a method of obtaining the density matrix representations in [2], where the pure states are particular senses of a word, and the general density matrix representations are weighted mixtures of its senses. The detection of relevant features for such an algorithm is another implementation issue that needs to be addressed.

The partial order I define on density matrices corresponds to the orthomodular lattice of the subspaces of a Hilbert space. Unfortunately, the negation defined through this logic does not resemble negation in natural language. Moreover, it can be argued that "good" can be used appropriately in any context "bad" can be used. In that sense, its inability to distinguish synonyms and antonyms is a weakness of this model, as it is of many other distributional models. It is simply not immediately obvious how the contexts of synonyms and antonyms differ.

Bibliography

[1] Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 23–32. Association for Computational Linguistics, 2012.

[2] Peter D Bruza and Richard Cole. Quantum logic of semantic space: An exploratory investigation of context effects in practical reasoning. *We Will Show Them: Essays in Honour of Dov Gabbay*, pages 339–361, 2005.

[3] Daoud Clarke. Context-theoretic semantics for natural language: an overview. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 112–119. Association for Computational Linguistics, 2009.

[4] Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*, 2010.

[5] James Richard Curran. From distributional to semantic similarity. 2004.

[6] Susan T Dumais. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230, 2004.

[7] Katrin Erk. Supporting inferences in semantic space: representing words as regions. In *Proceedings of the Eighth International Conference on Computational Semantics*, pages 104–115. Association for Computational Linguistics, 2009.

[8] Christiane Fellbaum. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010.

[9] JR Firth. Papers in general linguistics 1934–1951, 1957.

[10] Peter Gärdenfors. *Conceptual spaces: The geometry of thought*. MIT press, 2004.

[11] Maayan Geffet and Ido Dagan. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 107–114. Association for Computational Linguistics, 2005.

[12] Edward Grefenstette. Category-theoretic quantitative compositional distributional models of natural language semantics. *arXiv preprint arXiv:1311.1539*, 2013.

[13] Edward Grefenstette and Mehrnoosh Sadrzadeh. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404. Association for Computational Linguistics, 2011.

[14] Gregory Grefenstette. *Explorations in automatic thesaurus discovery*. Springer, 1994.

[15] Zellig S Harris. Distributional structure. *Word*, 1954.

[16] Aurélie Herbelot and Mohan Ganesalingam. Measuring semantic content in distributional vectors. In *ACL (2)*, pages 440–445, 2013.

[17] Chris Heunen and Jamie Vicary. *Introduction to categorical quantum mechanics*. Clarendon press, Oxford, 2014.

[18] Theo M. V. Janssen. Montague semantics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition, 2012.

[19] Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(04):359–389, 2010.

[20] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86, 1951.

[21] Joachim Lambek. Type grammars as pregroups. *Grammars*, 4(1):21–39, 2001.

[22] Lillian Lee. Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 25–32. Association for Computational Linguistics, 1999.

[23] Alessandro Lenci and Giulia Benotto. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 75–79. Association for Computational Linguistics, 2012.

[24] Dekang Lin. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304, 1998.

[25] Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208, 1996.

[26] Saunders Mac Lane. *Categories for the working mathematician*. springer, 1998.

[27] Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 279. Association for Computational Linguistics, 2004.

[28] Richard Montague. *English as a formal language*. Ed. di Comunità, 1970.

[29] Richard Montague. Universal grammar. *Theoria*, 36(3):373–398, 1970.

[30] Michael A Nielsen and Isaac L Chuang. *Quantum computation and quantum information*. Cambridge university press, 2010.

[31] Barbara Partee. Compositionality. *Varieties of formal semantics*, 3:281–311, 1984.

[32] Barbara Partee. Lexical semantics and compositionality. *An invitation to cognitive science: Language*, 1:311–360, 1995.

[33] Barbara H Partee. Formal semantics. In *Lectures at a workshop in Moscow. http://people. umass. edu/partee/RGGU_2005/RGGU05_formal_semantics. htm*, 2005.

[34] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 1*, pages 448–453. Morgan Kaufmann Publishers Inc., 1995.

[35] Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. Chasing hypernyms in vector spaces with entropy. *EACL 2014*, page 38, 2014.

[36] Hinrich Schütze. Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123, 1998.

[37] Peter Selinger. A survey of graphical languages for monoidal categories. In *New structures for physics*, pages 289–355. Springer, 2011.

[38] Kazuhide Sugawara, MasufiI Nishimura, Kolchi Toshioka, Masaaki Okochi, and Tovohisa Kaneko. Isolated word recognition using hidden markov models. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.*, volume 10, pages 1–4. IEEE, 1985.

[39] Hisaharu Umegaki. Conditional expectation in an operator algebra. iv. entropy and information. In *Kodai Mathematical Seminar Reports*, volume 14, pages 59–85. Tokyo Institute of Technology, Department of Mathematics, 1962.

[40] Salvatore Valenti, Francesca Neri, and Alessandro Cucchiarelli. An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, 2(1):319–330, 2003.

[41] Cornelis Joost Van Rijsbergen. *The geometry of information retrieval*, volume 157. Cambridge University Press Cambridge, 2004.

[42] Vlatko Vedral. The role of relative entropy in quantum information theory. *Reviews of Modern Physics*, 74(1):197, 2002.

[43] Julie Weeds, David Weir, and Diana McCarthy. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics*, page 1015. Association for Computational Linguistics, 2004.