Conversational Negation using Worldly Context in Compositional Distributional Semantics

Benjamin Rodatz
Computer Science,
University of Oxford
benjamin.rodatz
@cs.ox.ac.uk

Razin A. Shaikh
Mathematical Institute,
University of Oxford
razin.shaikh
@maths.ox.ac.uk

Lia Yeh
Quantum Group,
Computer Science,
University of Oxford
lia.yeh @cs.ox.ac.uk

All authors have contributed equally.

Abstract

We propose a framework to model an operational conversational negation by applying *worldly context* (prior knowledge) to logical negation in compositional distributional semantics. Given a word, our framework can create its negation that is similar to how humans perceive negation. The framework corrects logical negation to weight meanings closer in the entailment hierarchy more than meanings further apart. The proposed framework is flexible to accommodate different choices of logical negations, compositions, and worldly context generation. In particular, we propose and motivate a new logical negation using matrix inverse.

We validate the sensibility of our conversational negation framework by performing experiments, leveraging density matrices to encode graded entailment information. We conclude that the combination of subtraction negation ($\neg_{sub}$) and phaser in the basis of the negated word yields the highest Pearson correlation of 0.635 with human ratings.

1 Introduction

Negation is fundamental to every human language, marking a key difference from how other animals communicate (Horn, 1972). It enables us to express denial, contradiction, and other uniquely human aspects of language. As humans, we know that negation has an operational interpretation: if we know the meaning of *A*, we can infer the meaning of *not A*, without needing to see or hear *not A* explicitly in any context.

Formalizing an operational description of how humans interpret negation in natural language is a challenge of significance to the fields of linguistics, epistemology, and psychology. Kruszewski et al. (2016) notes that there is no straightforward negation operation that, when applied to the distributional semantics vector of a word, derives a negation of that word that captures our intuition. This work proposes and experimentally validates an operational framework for conversational negation in compositional distributional semantics.

In the field of distributional semantics, there have been developments in capturing the purely logical form of negation. Widdows and Peters (2003) introduce the idea of computing negation by mapping a vector to its orthogonal subspace; Lewis (2020) analogously model their logical negation for density matrices. However, logical negation alone is insufficient in expressing the nuances of negation in human language. Consider the sentences:

a) `This is not an apple; this is an orange.`

b) `This is not an apple; this is a paper.`

Sentence a) is more plausible in real life than sentence b). However, since apples and oranges share a lot in common, their vector or density matrix encodings would most likely not be orthogonal. Consequently, such a logical negation of apple would more likely indicate a paper than an orange.

Blunsom et al. (2013) propose that the encoding of a word should have a distinct "domain" and "value", and its negation should only affect the "value". In this way, *not blue* would still be in the domain of *color*. However, they do not provide any scalable way to generate such representation of "domain" and "value" from a corpus. We argue that this domain need not be encoded in the vector or density matrix itself. Instead, we propose a method to generate what we call *worldly context* directly from the word and its relationships to other words, computed a priori using worldly knowledge.

Furthermore, we want such conversational negation to generalize from words to sentences and to entire texts. DisCoCat (Coecke et al., 2010) provides a method to compose the meaning of words

to get the meaning of sentences and DisCoCirc (Coecke, 2020) extends this to propagate knowledge throughout the text. Therefore, we propose our conversational negation in the DisCoCirc formalism, putting our framework in a rich expanse of grammatical types and sentence structures. Focusing on the conversational negation of single words, we leave the interaction of conversational negation with grammatical structures for future work.

Section 2 introduces the necessary background. Section 3 discusses the logical negation using subtraction from the identity matrix from Lewis (2020), and proposes and justifies a second, new form of logical negation using matrix inverse. Section 4 introduces methods for context creation based on worldly knowledge. Section 5 presents the general framework for performing conversational negation of a word by combining logical negation with worldly context. Section 6 experimentally verifies the proposed framework, comparing each combination of different logical negations, compositions, bases, and worldly context generation. We end our discussion with an overview of future work.

2 Background

2.1 Conversational negation

Kruszewski et al. (2016) point out a long tradition in formal semantics, pragmatics and psycholinguistics which has argued that negation—in human conversation—is not simply a denial of information; it also indicates the truth of an *alternative* assertion. They call this alternative-licensing view of negation *conversational negation*.

Another view on negation states that the effect of negation is merely one of information denial (Evans et al., 1996). However, Prado and Noveck (2006) explain that even under this view, the search for alternatives could happen as a secondary effort for interpreting negation in the sentence.

The likelihood of different alternatives to a negated word inherently admits a grading (Oaksford, 2002; Kruszewski et al., 2016). For example, something that is not a *car* is more likely to be a *bus* than a *pen*. They argue that the most plausible alternatives are the ones that are applicable across many varied contexts; *car* can be replaced by *bus* in many contexts, but it requires an unusual context to sensibly replace *car* with *pen*.

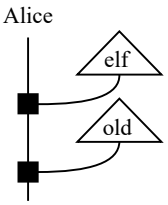


Figure 1: Graphical representation of meaning updating in DisCoCirc - read from top to bottom

2.2 Compositional semantics and DisCoCirc

Language comprehension depends on understanding the meaning of words as well as understanding how the words interact with each other in a sentence. While the former is an understanding of the definitions of words, the latter requires an understanding of grammar. Coecke et al. (2010) build on this intuition to propose DisCoCat, a compositional distributional model of meaning, making use of the diagrammatic calculus originally introduced for quantum computing (Abramsky and Coecke, 2004). In Coecke (2020), this model was extended to DisCoCirc which generalized DisCoCat from modeling individual sentences to entire texts. In DisCoCirc, the two sentences

```
Alice is an elf.
Alice is old.
```

are viewed as two processes updating the state of Alice, about whom, at the beginning of the text, the reader knows nothing. Graphically this can be displayed as shown in Figure 1. The wire labeled by *Alice* represents the knowledge we have about Alice at any point in time. It is first updated by the fact that she is an elf and subsequently updated by the fact that she is old. We use a black square to represent a general meaning-update operation, which can be one of a variety of operators we discuss in the next section. DisCoCirc allows for more grammatically complex sentence and text structures not investigated in this work.

DisCoCirc allows for various ways of representing meaning such as vector spaces (Coecke et al., 2010; Grefenstette and Sadrzadeh, 2011), conceptual spaces (Bolt et al., 2017), and density matrices (Balkir et al., 2016; Lewis, 2019). A density matrix is a complex matrix, which is equal to its own conjugate transpose (Hermitian) and has non-negative eigenvalues (positive semidefinite). They can be viewed as an extension of vector spaces to allow for encoding lexical entailment structure (see Section 2.4), a property for which they were selected

as the model of meaning for this paper.

2.3 Compositions for meaning update

We present four compositions for meaning update:

$$\text{spider}(\mathsf{A}, \mathsf{B}) := U_s(\mathsf{A} \otimes \mathsf{B})U_s^\dagger \quad (1)$$

- $U_s = \sum_i |i\rangle \langle ii|$ where $\{|i\rangle\}_i$ is B's eigenbasis
- non-linear AND in Coecke (2020)

$$\text{fuzz}(\mathsf{A}, \mathsf{B}) := \sum_i x_i P_i \circ \mathsf{A} \circ P_i \quad (2)$$

- $\mathsf{B} = \sum_i x_i P_i$
- in Coecke and Meichanetzidis (2020)
- Kmult in Lewis (2020)

$$\text{phaser}(\mathsf{A}, \mathsf{B}) := \mathsf{B}^{\frac{1}{2}}\mathsf{A}\mathsf{B}^{\frac{1}{2}} \quad (3)$$

- $\mathsf{B} = \sum_i x_i^2 P_i$ where $\mathsf{B}^{\frac{1}{2}} = \sum_i x_i P_i$
- in Coecke and Meichanetzidis (2020)
- Bmult in Lewis (2020)
- corresponds to quantum Bayesian update (van de Wetering, 2018)

$$\text{diag}(\mathsf{A}, \mathsf{B}) := dg(\mathsf{A}) \circ dg(\mathsf{B}) \quad (4)$$

- a Compr from De las Cuevas et al. (2020): lifts verbs and adjectives to completely positive maps matching their grammatical type

where A and B are density matrices, x_i is a real scalar between 0 and 1, P_i's are projectors, and the function *dg* sets all off-diagonal matrix elements to 0 giving a diagonal matrix.

Of the many Compr variants (De las Cuevas et al., 2020), we only consider *diag* and *mult* (elementwise matrix multiplication, which is an instance of *spider*) as candidates for composition. All other variants are scalar multiples of one input, the identity wire, or a maximally mixed state; therefore we do not consider them as they discard too much information about the inputs.

For *spider*, *fuzz*, and *phaser*, choosing the basis of the composition determines the basis the resulting density matrix takes on, and its meaning is interpreted in (Coecke and Meichanetzidis, 2020).

2.4 Lexical entailment via hyponymies

A word w_A is a hyponym of w_B if w_A is a type of w_B; then, w_B is a hypernym of w_A. For example, *dog* is a hyponym of *animal*, and *animal* is a hypernym of *dog*. Where there is a meaning relation between two words, there exists an entailment relation between two sentences containing those words. Measures to quantify these relations ought to be *graded*, as one would expect some entailment relations to be weaker than others. Furthermore, such measures should be *asymmetric* (a bee is an insect, but an insect is not necessarily a bee) and *pseudo-transitive* (a t-shirt is a shirt, a shirt can be formal, but a t-shirt is usually not formal).

One of the limitations of the vector space model of NLP is that it does not admit a natural non-trivial graded entailment structure (Balkir et al., 2016; Coecke, 2020). Bankova et al. (2019) utilize the richer setting of density matrices to define a measure called k-hyponymy, generalizing the Löwner order to have a grading for positive operators, satisfying the above three properties. They further lift from entailment between words to between two sentences of the same grammatical structure, using compositional semantics, and prove a lower bound on this entailment between sentences.

The k-hyponymy (k_{hyp}) between density matrices A and B is the maximum k such that

$$\mathsf{A} \sqsubseteq_k \mathsf{B} \iff \mathsf{B} - k\,\mathsf{A} \text{ is a positive operator} \quad (5)$$

where k is between 0 (no entailment) and 1 (full entailment).

Van de Wetering (2018) finds that the crisp Löwner ordering ($k_{\mathsf{hyp}} = 1$) is trivial when operators are normalized to trace 1. On the other hand, they enumerate highly desirable properties of the Löwner order when normalized to highest eigenvalue 1. In particular, the maximally mixed state is the bottom element; all pure states are maximal; and the ordering is preserved under any linear trace-preserving isometry (including unitaries), convex mixture, and the tensor product. In our experiments, we leverage these ordering properties following Lewis (2020)'s convention of normalizing operators to highest eigenvalue ≤ 1.

According to Bankova et al. (2019, Theorem 2), when $supp(\mathsf{A}) \subseteq supp(\mathsf{B})$, k_{hyp} is given by $1/\gamma$, where γ is the maximum eigenvalue of $\mathsf{B}^+\mathsf{A}$. Here B^+ denotes the Moore-Penrose inverse of B, which we refer to in the next section as support inverse. If $supp(\mathsf{A}) \not\subseteq supp(\mathsf{B})$, k_{hyp} is 0. This means that

k_{hyp} admits a grading, but is not robust to errors. In our experiments, to circumvent this issue of almost all of our calculated k_{hyp} being 0, we employ a generalized form of k_{hyp} equivalent to as originally defined in Bankova et al. (2019, Theorem 2), less checking whether $supp(\mathsf{A}) \subseteq supp(\mathsf{B})$.

To propose more robust measures, Lewis (2019) says A entails B with the error term E if there exists a D such that:

$$\mathsf{A} + \mathsf{D} = \mathsf{B} + \mathsf{E} \tag{6}$$

to define the following two entailment measures

$$k_{\mathsf{BA}} = \frac{\sum_i \lambda_i}{\sum_i |\lambda_i|} = \frac{\mathsf{Trace}(\mathsf{D} - \mathsf{E})}{\mathsf{Trace}(\mathsf{D} + \mathsf{E})} \tag{7}$$

$$k_{\mathsf{E}} = 1 - \frac{\|\mathsf{E}\|}{\|\mathsf{A}\|} \tag{8}$$

where the λ_i's are the eigenvalues of $\mathsf{B} - \mathsf{A}$. In Equations 7 and 8, the error term E satisfying Equation 6 is constructed by taking the diagonalization of $\mathsf{B} - \mathsf{A}$, setting all positive eigenvalues to zero, and changing the sign of all negative eigenvalues. k_{BA} ranges from -1 to 1, and k_{E} ranges from 0 to 1.

According to De las Cuevas et al. (2020), *diag*, *mult*, and *spider* preserve crisp Löwner order:

$$\mathsf{A}_1 \sqsubseteq \mathsf{B}_1, \mathsf{A}_2 \sqsubseteq \mathsf{B}_2 \iff \mathsf{A}_1 \between \mathsf{A}_2 \sqsubseteq \mathsf{B}_1 \between \mathsf{B}_2 \tag{9}$$

Fuzz and *phaser* do not satisfy Equation 9.

3 Logical negations

To construct conversational negation, we must first define a key ingredient – logical negation, denoted by $\neg$. The logical negation of a density matrix is a unary function that yields another density matrix.

The most important property of a logical negation is that it must interact well with hyponymy. Ideally, the interpretation of the contrapositive of an entailment must be sensible:

$$\mathsf{A} \sqsubseteq \mathsf{B} \iff \neg\mathsf{B} \sqsubseteq \neg\mathsf{A} \tag{10}$$

A weakened notion arises from allowing varying degrees of entailment:

$$\mathsf{A} \sqsubseteq_k \mathsf{B} \iff \neg\mathsf{B} \sqsubseteq_{k'} \neg\mathsf{A} \tag{11}$$

where $k = k'$ in the ideal case.

Equation 11 necessitates any candidate of logical negation to be *order-reversing*. However, van de Wetering (2018) proved that all unitary operations preserve Löwner order. Therefore, no quantum gates can reverse Löwner order, and the search for a logical negation compatible with quantum natural language processing (Coecke et al., 2020) (originally formulated in the category of **CPM(FHilb)** (Piedeleu et al., 2015)) remains an open question.

We now discuss two candidates for logical negation that have desirable properties and interaction with the hyponymies presented in Section 2.4.

3.1 Subtraction from identity negation

Lewis (2020) introduces a candidate logical negation which preserves positivity of density matrix X:

$$\neg_{sub}\mathsf{X} := \mathbb{I} - \mathsf{X} \tag{12}$$

In the case where X is a pure state, it maps X to the subspace orthogonal to it, as the identity matrix $\mathbb{I}$ is the sum of orthonormal projectors. This logical negation satisfies Equation 10 for the crisp Löwner order. It satisfies Equation 11 with $k = k'$ for k_{BA}, but not for k_{hyp} or k_{E}.

3.2 Matrix inverse negation

We introduce a new candidate for logical negation, the *matrix inverse*. This reverses Löwner order, i.e. satisfies Equation 11 with $k = k'$ (see Corollary 1 in Appendix). It additionally satisfies Equation 11 with $k = k'$ for k_{BA} if both density operators have same eigenbases (see Theorem 2 in Appendix).

As the matrix inverse of a non-invertible matrix is undefined, we define a logical negation from two generalizations of the matrix inverse acting upon the support and kernel subspaces, respectively.

Definition 1. *For any density matrix X with spectral decomposition $\mathsf{X} = \sum_i \lambda_i |i\rangle \langle i|$,*

$$\neg_{supp}\mathsf{X} := \sum_i \begin{cases} \frac{1}{\lambda_i} |i\rangle \langle i|, & \text{if } \lambda_i > 0 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

Definition 1 is the Moore-Penrose generalized matrix inverse and is equal to the matrix inverse when the kernel is empty. It has the property that Equation 11 with $k = k'$ is satisfied for k_{hyp} when $rank(\mathsf{A}) = rank(\mathsf{B})$ (see Theorem 1 in Appendix). We call it the *support inverse*, to contrast with what we call the *kernel inverse*:

Definition 2. *For any non-invertible density matrix X with spectral decomposition $\mathsf{X} = \sum_i \lambda_i |i\rangle \langle i|$,*

$$\neg_{ker}\mathsf{X} := \sum_i \begin{cases} 1 |i\rangle \langle i|, & \text{if } \lambda_i = 0 \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

The kernel inverse is the limit of matrix regularization by spectral filtering (i.e. setting all zero eigenvalues to an infinitesimal positive eigenvalue), then inverting the matrix and normalizing to highest eigenvalue 1. Its application discards all information about the eigenspectrum of the original matrix. Therefore, applying the kernel inverse twice results in a maximally mixed state over the support of the original matrix. Operationally speaking, $\neg_{ker}$ and $\neg_{sub}$ act upon the kernel of the original matrix identically.

We can think conceptually of a negated word as containing elements both "near" (in support) and "far" (in kernel) from the original word. Therefore, a logical negation should encompass nonzero values in the original matrix's support and in its kernel; it is up to conversational negation to then weight the values in the logical negation according to their contextual relevance.

On their own, neither the support inverse nor the kernel inverse are sensible candidates for logical negation. A convex mixture of the two, which we call *matrix inverse* and denote with $\neg_{inv}$, spans both support and kernel of the original matrix. In our experiments we weight support and kernel equally, but other weightings could be considered, for instance to take into account a noise floor or enforce the naively unsatisfied property that twice application is the identity operation.

When composing a density matrix X with $\neg_{inv}\mathsf{X}$ or $\neg_{supp}\mathsf{X}$ via *spider*, *fuzz*, or *phaser*, the resulting density matrix has the desired property of being a maximally mixed state on the support with zeroes on the kernel (see Theorem 3 and Corollary 2 in Appendix). In other words, this operation is the fastest "quantum (Bayesian, in the case of *phaser*) update" from a density matrix to the state encoding no information other than partitioning support and kernel subspaces. Interpreting composition as logical AND, this corresponds to the contradiction that a proposition (restricted to the support subspace) cannot simultaneously be true and not true.

3.3 Normalization

$\neg_{sub}$, $\neg_{supp}$, and $\neg_{inv}$ preserve eigenvectors (up to uniqueness for eigenvalues with multiplicity > 1). We ignore normalization for logical negation because in our conversational negation framework, which we introduce in Section 5, we can always normalize to largest eigenvalue ≤ 1 after the composition operation.

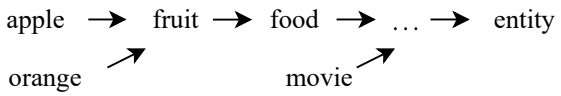


Figure 2: Example of hyponymy structure as can be found in entailment hierarchies

4 Context determination

Negation is intrinsically dependent on context. Context can be derived from two sources: 1) knowledge gained throughout the sentence or the text (textual context), and 2) worldly knowledge from experts or data such as a corpus (worldly context). While textual context depends on the specific text being analyzed, worldly context can be computed a priori. In this section, we introduce worldly context and propose two methods of computing it.

4.1 Worldly context

Worldly knowledge is a certain understanding of the world that most users of a language intuitively possess. We want to capture this worldly knowledge to provide a context for negation that is not explicit in the text. In this section, we propose two methods of generating a worldly context: 1) knowledge encoded in an entailment hierarchy such as WordNet, and 2) generalizing the ideas of the first method to context derivation from the entailment information encoded in density matrices.

4.1.1 Context from an entailment hierarchy

We consider an entailment hierarchy for words that leads to relations such as in Figure 2, where a directed edge can be understood as a hyponym relation. Such relational hierarchy can be obtained from human curated database like WordNet (Fellbaum, 1998) or using unsupervised methods such as Hearst patterns (Hearst, 1992; Roller et al., 2018).

We can use such a hierarchy of hyponyms to generate worldly context, as words usually appear in the implicit context of their hypernyms; for example, *apple* is usually thought of as a *fruit*. Now, to calculate the worldly context for the word *apple*, we take a weighted sum of the hypernyms of *apple*, with more direct hypernyms such as *fruit* weighted higher than more distant hypernyms such as *entity*. This corresponds to the idea that when we talk in the context of *apple*, we are more likely to talk about an *orange* (hyponym of *fruit*) than a *movie* (hyponym of *entity*). Hence, for a word w

with hypernyms $h_1, \ldots, h_n$ ordered from closest to furthest, we define the worldly context $\mathtt{wc}_w$ as:

$$[\![\mathtt{wc}_w]\!] := \sum_i p_i [\![h_i]\!] \tag{15}$$

where $p_i \geq p_{i+1}$ for all i.

For this approach, we assume that the density matrix of the word is a mixture containing its hyponyms; i.e. the density matrix of *fruit* is a mixture of all fruits such as *apple*, *orange* and *pears*.

4.1.2 Context using entailment encoded in the density matrices

As explained in Section 2.4, density matrix representation of words can be used to encode the information about entailment between words. Furthermore, this entailment can be graded; for example, *fruit* would entail *dessert* with a high degree, but not necessarily by 1. Such graded entailment is not captured in the human curated WordNet database. Although there have been proposals to extend WordNet (Boyd-Graber et al., 2006; Ahsaee et al., 2014), such semantic networks are not yet available.

We generalize the idea of entailment hierarchy by considering a directed weighted graph where each node is a word and the edges indicate how much one word entails the other. Once we have the density matrices for words generated from corpus data, we can build this graph by calculating the graded hyponymies (see Section 2.4) among the words, thereby extracting the knowledge gained from the corpus encoded in the density matrices, without requiring human narration.

Consider words x and y where $x \sqsubseteq_p y$ and $y \sqsubseteq_q x$. In the ideal case, there are three possibilities: 1) x and y are not related (both p and q are small), 2) one is a type of the other (one of p and q is large), or 3) they are very similar (both p and q are large). Hence, we need to consider both p and q when we generate the worldly context. To obtain the worldly context for a word w, we consider all nodes (words) connected to w along with their weightings. If $p_1, \ldots, p_n$ and $q_1, \ldots, q_n$ are the weights of the edges from w to words $h_1, \ldots, h_n$, then worldly context $\mathtt{wc}_w$ is given by

$$[\![\mathtt{wc}_w]\!] := \sum_i f(p_i, q_i) [\![h_i]\!] \tag{16}$$

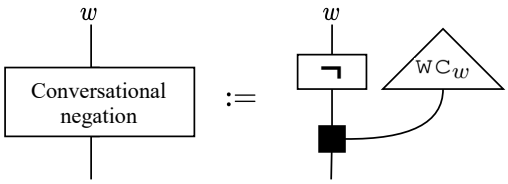
where f is some function of weights p_i and q_i.

5 Conversational negation in DisCoCirc

5.1 A framework for conversational negation

In this section, we present a framework to obtain conversational negation by composing logical negation with worldly context. As discussed in Section 2.1, negation—when used in conversation—can be viewed as not just a complement of the original word, but as also suggesting an *alternative* claim. Therefore, to obtain conversational negation, we need to adapt the logical negation to take into account the worldly context of the negated word.

In DisCoCirc (see Section 2.2), words are wires, and sentences are processes that update meaning of the words. Similarly, we view *conversational negation* as a process that updates the meaning of the words. We propose the general framework for conversational negation by defining it to be the logical negation of the word, updated through composition with the worldly context evoked by that word:



(17)

The framework presented here is general; i.e. it does not restrict the choice of logical negation, worldly context or composition. The main steps of conversational negation are:

1. Calculate the logical negation $\neg([\![w]\!])$.
2. Compute the worldly context $[\![\mathtt{wc}_w]\!]$.
3. Update the meaning of $\neg([\![w]\!])$ by composing with $[\![\mathtt{wc}_w]\!]$ to obtain $\neg([\![w]\!]) \curvearrowright [\![\mathtt{wc}_w]\!]$.

Further meaning updates can be applied to the output of conversational negation using compositional semantics as required from the structure of the text, although we do not investigate this in the current work.

5.2 See it in action

We present a toy example to develop intuition of how meaning provided by worldly context interacts with logical negation and composition to derive conversational negation. Suppose $\{apple,\ orange,\ fig,\ movie\}$ are pure states forming an orthonormal basis (ONB). In practice ONBs are far larger, but this example suffices to illustrate how the conversational negation accounts for which states are relevant. We take $\neg_{sub}$ as the

choice of negation and *spider* in this ONB as the choice of composition.

Now, consider the sentence:

`This is not an apple.`

Although in reality the worldly context of $apple$ encompasses more than just $fruit$, for ease of understanding, assume the worldly context of apple is $[\![\mathtt{w}\subset_{apple}]\!] = [\![fruit]\!]$, given by

$$[\![fruit]\!] = \frac{1}{2}[\![apple]\!] + \frac{1}{3}[\![orange]\!] + \frac{1}{6}[\![fig]\!]$$

Applying $\neg_{sub}([\![apple]\!]) = \mathbb{I} - [\![apple]\!]$, we get

$$\neg_{sub}([\![apple]\!]) = [\![orange]\!] + [\![fig]\!] + [\![movie]\!]$$

Finally, to obtain conversational negation, logical negation is endowed with meaning through the application of worldly context.

$$\neg_{sub}([\![apple]\!]) \curlywedge [\![fruit]\!] = \frac{1}{3}[\![orange]\!] + \frac{1}{6}[\![fig]\!]$$

This conversational negation example not only yields all *fruits* which are not *apples*, but also preserves the proportions of the non-apple fruits.

6 Experiments

To validate the proposed framework, we perform experiments on the data set of alternative plausibility ratings created by Kruszewski et al. (2016)[1]. In their paper, Kruszewski et al. (2016) predict plausibility scores for word pairs consisting of a negated word and its alternative using various methods to compare the similarity of the words. While achieving a high correlation with human intuition, they do not provide an operation to model the outcome of a conversational negation. Through the experiments, we test whether our operational conversational negation still has correlation with human intuition.

6.1 Data

The Kruszewski et al. (2016) data set consists of word pairs containing a noun to be negated and an alternative noun, along with a plausibility rating. We will denote the word pairs as (w_N, w_A). The authors transform these word pairs into simple sentences of the form: *This is not a* w_N*, it is a* w_A (e.g. This is not a radio, it is a dad.). These sentences are then rated by human participants on how plausible they are to appear in a natural conversation.

To build these word pairs, Kruszewski et al. (2016) randomly picked 50 common nouns as w_N and paired them with alternatives that have various relations to w_N. Then using a crowd-sourcing service, they asked the human participants to judge the plausibility of each sentence. The participants were told to rate each sentence on a scale of 1 to 5.

6.2 Methodology

We build density matrices from 50 dimensional GloVe (Pennington et al., 2014) vectors using the method described in Lewis (2019). Then for each word pair (w_N, w_A) in the data set, we use various combinations of operations to perform conversational negation on the density matrix of w_N and calculate similarity with the density matrix of w_A.

For conversational negation, we experiment with different combinations of logical negations, composition operations and worldly context. We use two types of logical negations: $\neg_{sub}$ and $\neg_{inv}$. For composition, we use *spider*, *fuzz*, *phaser*, *mult* and *diag*. With *spider*, *fuzz* and *phaser*, we perform experiments in two choices of basis: 'w', the basis of $\neg([\![w_N]\!])$, and 'c', the basis of $[\![\mathtt{w}\subset_{w_N}]\!]$. We use worldly context generated from the WordNet entailment hierarchy as per Section 4.1.1; we experiment with different methods to calculate the weights p_i along the hypernym path.

To find plausibility ratings, we calculate hyponymies k_{hyp}, k_{E} and k_{BA}, as well as *trace similarity* (the density operator analog of cosine similarity for vectors), between the density matrix of the conversational negation of w_N and $[\![w_A]\!]$. Note that in our experiments, unlike in the originally proposed formulation of k_{hyp}, we generalize k_{hyp} to not be 0 when $supp(\mathsf{A}) \not\subseteq supp(\mathsf{B})$, as described in Section 2.4. We calculate entailment in both directions for k_{E} and k_{hyp}, which are asymmetric. The entailment from w_N to w_A is denoted k_{E1} and k_{hyp1} while the entailment from w_A to w_N is denoted k_{E2} and k_{hyp2}. Finally, we calculate the Pearson correlation between our plausibility ratings and the mean human plausibility ratings from Kruszewski et al. (2016).

6.3 Results

Our experiments revealed that the best conversational negation is obtained by choosing $\neg_{sub}$ with *phaser* in the basis 'w'. We achieve 0.635 correlation of the *trace similarity* plausibility rating with

[1]The data set is available at http://marcobaroni.org/PublicData/alternatives_dataset.zip

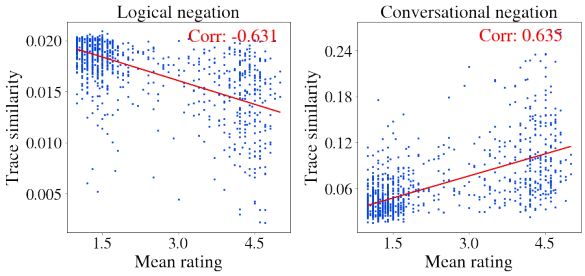


Figure 3: Correlation of logical (left) and conversational negation (right) with mean human rating

Logical negation	Composition	k_{E1}	k_{E2}	k_{hyp1}	k_{hyp2}	k_{BA}	trace
$\neg_{sub}$	$spider_c$	-0.152	0.068	0.287	0.357	0.241	0.355
	$spider_w$	-0.181	-0.176	0.282	0.217	0.243	-0.154
	$phaser_c$	-0.270	-0.279	0.305	0.153	0.256	-0.235
	$phaser_w$	0.432	0.602	0.289	0.495	0.311	0.635
	$fuzz_c$	-0.226	-0.064	0.298	0.175	0.246	0.449
	$fuzz_w$	-0.252	-0.125	0.304	0.191	0.259	0.047
$\neg_{inv}$	$spider_c$	0.197	0.455	0.263	0.419	0.261	0.455
	$spider_w$	0.006	-0.034	0.273	0.112	0.163	0.111
	$phaser_c$	-0.258	-0.129	0.285	0.139	0.183	-0.037
	$phaser_w$	0.279	0.432	0.285	0.285	0.241	0.519
	$fuzz_c$	-0.212	-0.050	0.296	0.034	0.188	0.135
	$fuzz_w$	-0.261	-0.070	0.299	0.180	0.232	0.033

Figure 4: Correlation of various conversational negations with mean plausibility ratings of human participants. Correlations above 0.4 are highlighted in green.

the human ratings, as shown in Figure 3 (right).

On the other hand, Figure 3 (left) shows *trace similarity* of $\neg_{sub}$ without applying any context. We observe that simply performing logical negation yields a negative correlation with human plausibility ratings. This is because logical negation gives us a density matrix furthest from the original word, going against the observation of Kruszewski et al. (2016) that an alternative to a negated word appears in similar contexts to it. Figure 3 (right) shows the results of combining this logical negation with worldly context to obtain meaning that positively correlates with how humans think of negation in conversation.

We tested many combinations for conversational negation enumerated in Section 6.2. The correlation between plausibility ratings for our conversational negation and the mean human plausibility rating is shown in Figure 4. We left out *mult* and *diag* from the table as they did not achieve any correlation above 0.3. Now, we will explore each variable of our experiments individually in the next sections.

6.3.1 Logical negation

We tested $\neg_{sub}$ and $\neg_{inv}$ logical negations. We found that the conversational negations built from $\neg_{sub}$ negation usually had a higher correlation with human plausibility ratings, with the highest being 0.635 as shown in Figures 3 and 4. One exception to this is when the $\neg_{inv}$ is combined with *spider* in the basis 'c', for which we get the correlation of 0.455 for both *trace similarity* and k_{E2}.

6.3.2 Composition

We investigated five kinds of composition operations: *spider*, *fuzz*, *phaser*, *mult*, and *diag*. We found that the results using *mult* and *diag* do not have any statistically significant correlation ($<$0.3) with human plausibility rating. On the other hand, *phaser* (in the basis 'w') has the highest correlation. It performs well with both logical negations. Plausibility ratings for *phaser* with $\neg_{sub}$ negation measured using k_{E2} and *trace similarity* has correlations of 0.602 and 0.635 respectively. *Spider* and *fuzz* have statistically relevant correlation for a few cases but never more than 0.5.

6.3.3 Basis

Spider, *fuzz*, and *phaser* necessitate a choice of basis for applying the worldly context in the conversational negation. We can interpret this choice as determining which input density matrix sets the eigenbasis of the output, and which modifies the other's spectrum. We found that *phaser* paired with the basis 'w' (the basis of the logically negated word) performs better than the basis 'c' (the basis of the worldly context) across both negations for most plausibility metrics. This lines up with our intuition that applying worldly context updates the eigenspectrum of $\neg([\![w_N]\!])$, leveraging worldly knowledge to increase/decrease the weights of more/less contextually relevant values of the logical negation of w_N. However, a notable exception to this reasoning is our result that for *spider* paired with $\neg_{inv}$, basis 'c' has statistically significant correlations with human ratings, while basis 'w' does not.

6.3.4 Worldly context

For these experiments, we create worldly context based on the hypernym paths provided by WordNet. As explained in Section 4.1.1, we need $p_i \geq p_{i+1}$ in Equation 15 for the more direct hypernyms to be more important than more distant hypernyms. Hence, we tried multiple monotonically decreasing functions for the weights $\{p_i\}_i$ of the hypernyms.

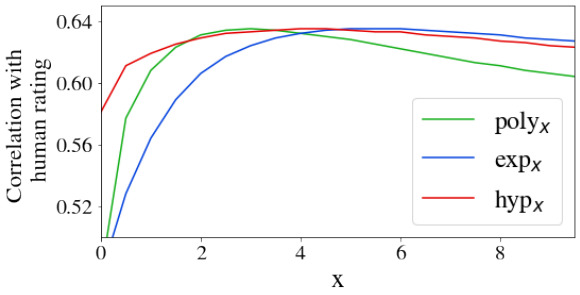


Figure 5: Correlation of results of different context functions with human rating

For a word w with n hypernyms $h_1, ..., h_n$ ordered from closest to furthest, we define the following functions to calculate p_i.

$$\texttt{poly}_x(i) \coloneqq (n-i)^x \tag{18}$$

$$\texttt{exp}_x(i) \coloneqq (1+\frac{x}{10})^{(n-i)} \tag{19}$$

$$\texttt{hyp}_x(i) \coloneqq (n-i)^{\frac{x}{2}} k_{\mathsf{E}}(w, h_i) \tag{20}$$

Figure 5 shows on the y-axis the correlation of the human rating with the plausibility rating (*trace*) of our best conversational negation (*phaser* with $\neg_{sub}$ in the basis 'w') and the parameters of context functions on the x-axis. We observe that all three context functions achieve a maximal correlation of 0.635, therefore being equally good. All functions eventually drop in correlation as the value of x increases, showing that having the context too close to the word does not yield optimal results either. One important observation is that at $x = 0$, $\texttt{hyp}_x(i) = k_{\mathsf{E}}(w, h_i)$ still performs well with a correlation of 0.581, despite not taking the WordNet hypernym distance into account. This is an evidence for the potential of the context creation based on density matrix entailment proposed in Section 4.1.2.

6.3.5 Plausibility rating measures

On top of calculating the conversational negation, the experiments call for comparing the results of the conversational negation with w_A to give plausibility ratings. We compare the hyponymies k_{E}, k_{hyp}, and k_{BA}, as well as *trace similarity*. The results show that *trace similarity* and k_{E2} interact most sensibly with our conversational negation, attaining 0.635 and 0.602 correlation with mean human ratings respectively. For the asymmetric measures k_{E} and k_{hyp}, computing the entailment from w_A to the conversational negation of w_N performed better than the other direction. For all similarity measures (except k_{hyp1}), $\neg_{sub}$ paired with *phaser* in the basis 'w' performs the best.

7 Future work

The framework presented in this paper shows promising results for conversational negation in compositional distributional semantics. Given its modular design, additional work should be done exploring more kinds of logical negations, compositions and worldly contexts, as well as situations for which certain combinations are optimal. Since creating worldly context—as presented in this paper—is a new concept in the area of DisCoCirc, it leaves the most room for further exploration. In particular, our framework does not handle how to disambiguate different meanings of the same word; for example, the worldly context of the word *apple* should be different for the fruit *apple* versus the technology company *apple*.

Our conversational negation framework currently does not model a different kind of negation where the suggested alternative is an antonym rather than just any other word that appears in similar contexts. For instance, the sentence *Alice is not happy* suggests that Alice is *sad*—an antonym of *happy*—rather than *cheerful*, even though *cheerful* might appear in similar contexts as *happy*. We would like to extend the conversational negation framework to account for this.

We would like to implement the context generation method presented in Section 4.1.2 and test on the current experimental setup.[2] To further validate the framework, more data sets should be collected and evaluated on to explore, for each type of relation between words, what construction of conversational negation yields sensible plausibility ratings.

For the conversational negation to be fully applicable in the context of compositional distributional semantics, further theoretical work is required to generalize the model from negation of individual nouns to negation of other grammatical classes and complex sentences. Furthermore, we would like to analyze the interplay between conversational negation, textual context, and evolving meanings. Lastly, the interaction of conversational negation with logical connectives and quantifiers leaves open questions to explore.

[2]The code is available upon request.

Acknowledgements

We would like to give special thanks to Martha Lewis for the insightful conversation and for sharing her code for generating density matrices. We appreciate the guidance of Bob Coecke in introducing us to the field of compositional distributional semantics for natural language processing. We thank John van de Wetering for informative discussion about ordering density matrices. We thank the anonymous reviewers for their helpful feedback. Lia Yeh gratefully acknowledges funding from the Oxford-Basil Reeve Graduate Scholarship at Oriel College in partnership with the Clarendon Fund.

References

Samson Abramsky and Bob Coecke. 2004. A categorical semantics of quantum protocols. In *Proceedings of the 19th Annual IEEE Symposium on Logic in Computer Science, 2004.*, pages 415–425.

Mostafa Ghazizadeh Ahsaee, Mahmoud Naghibzadeh, and S Ehsan Yasrebi Naeini. 2014. Semantic similarity assessment of words using weighted wordnet. *International Journal of Machine Learning and Cybernetics*, 5(3):479–490.

Jerzy K. Baksalary, Friedrich Pukelsheim, and George P.H. Styan. 1989. Some properties of matrix partial orderings. *Linear Algebra and its Applications*, 119:57–85.

Esma Balkir, Mehrnoosh Sadrzadeh, and Bob Coecke. 2016. Distributional sentence entailment using density matrices. In *Topics in Theoretical Computer Science*, pages 1–22, Cham. Springer International Publishing.

Dea Bankova, Bob Coecke, Martha Lewis, and Dan Marsden. 2019. Graded hyponymy for compositional distributional semantics. *Journal of Language Modelling*, 6(2):225–260.

Phil Blunsom, Edward Grefenstette, and Karl Moritz Hermann. 2013. "not not bad" is not "bad": A distributional account of negation. In *Proceedings of the 2013 Workshop on Continuous Vector Space Models and their Compositionality*.

Joe Bolt, Bob Coecke, Fabrizio Genovese, Martha Lewis, Dan Marsden, and Robin Piedeleu. 2017. Interacting conceptual spaces I : Grammatical composition of concepts. *CoRR*, abs/1703.08314.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to wordnet. In *Proceedings of the third international WordNet conference*, pages 29–36. Citeseer.

Bob Coecke. 2020. The mathematics of text structure.

Bob Coecke, Giovanni de Felice, Konstantinos Meichanetzidis, and Alexis Toumi. 2020. Foundations for near-term quantum natural language processing. *ArXiv*, abs/2012.03755.

Bob Coecke and Konstantinos Meichanetzidis. 2020. Meaning updating of density matrices. *FLAP*, 7:745–770.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *Lambek Festschrift Linguistic Analysis*, 36.

Gemma De las Cuevas, Andreas Klinger, Martha Lewis, and Tim Netzer. 2020. Cats climb entails mammals move: preserving hyponymy in compositional distributional semantics. In *Proceedings of SEMSPACE 2020*.

Jonathan St BT Evans, John Clibbens, and Benjamin Rood. 1996. The role of implicit and explicit negation in conditional reasoning bias. *Journal of Memory and Language*, 35(3):392–409.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database.

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

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Coling 1992 volume 2: The 15th international conference on computational linguistics*.

Laurence Horn. 1972. On the semantic properties of logical operators in english. *Unpublished Ph.D. dissertation*.

Germán Kruszewski, Denis Paperno, Raffaella Bernardi, and Marco Baroni. 2016. There is no logical negation here, but there are alternatives: Modeling conversational negation with distributional semantics. *Computational Linguistics*, 42(4):637–660.

Martha Lewis. 2019. Compositional hyponymy with positive operators. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 638–647, Varna, Bulgaria. INCOMA Ltd.

Martha Lewis. 2020. Towards logical negation for compositional distributional semantics. *IfCoLoG Journal of Logics and their Applications*, 7(3).

Mike Oaksford. 2002. Contrast classes and matching bias as explanations of the effects of negation on conditional reasoning. *Thinking & Reasoning*, 8(2):135–151.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Robin Piedeleu, Dimitri Kartsaklis, Bob Coecke, and Mehrnoosh Sadrzadeh. 2015. Open system categorical quantum semantics in natural language processing.

Jérôme Prado and Ira A. Noveck. 2006. How reaction times can elucidate matching effects and the processing of negation. *Thinking and Reasoning*, 12(3).

Stephen Roller, Douwe Kiela, and Maximilian Nickel. 2018. Hearst patterns revisited: Automatic hypernym detection from large text corpora. *arXiv preprint arXiv:1806.03191*.

John van de Wetering. 2018. Ordering quantum states and channels based on positive bayesian evidence. *Journal of Mathematical Physics*, 59(10):102201.

Dominic Widdows and Stanley Peters. 2003. Word vectors and quantum logic: Experiments with negation and disjunction. *Mathematics of language*, 8(141-154).

A Proofs

A.1 Support inverse reverses k-hyponymy

Theorem 1. *For two density matrices A and B, k-hyponymy is reversed by support inverse when $rank(A) = rank(B)$:*

$$A \sqsubseteq_k B \iff \neg_{supp} B \sqsubseteq_k \neg_{supp} A \tag{21}$$

Proof. From (Baksalary et al., 1989), $\neg_{supp}$ reverses Löwner order when $rank(\mathsf{A}) = rank(\mathsf{B})$:

$$\mathsf{A} \sqsubseteq \mathsf{B} \iff \neg_{supp}\mathsf{B} \sqsubseteq \neg_{supp}\mathsf{A} \tag{22}$$

Thus, letting "≥ 0" denote the operator is positive:

$$\begin{aligned}
\mathsf{A} \sqsubseteq_k \mathsf{B} &\iff \mathsf{B} - k\mathsf{A} \geq 0 && (23)\\
&\iff (k\mathsf{A})^{-1} - \mathsf{B}^{-1} \geq 0 && (24)\\
&\iff \frac{1}{k}\mathsf{A}^{-1} - \mathsf{B}^{-1} \geq 0 && (25)\\
&\iff \mathsf{A}^{-1} - k\mathsf{B}^{-1} \geq 0 && (26)\\
&\iff \mathsf{B}^{-1} \sqsubseteq_k \mathsf{A}^{-1} && (27)
\end{aligned}$$

using Equations 5 and 22 from Equation 23 to 24. □

Corollary 1. *For two invertible density matrices A and B, k-hyponymy is reversed by matrix inverse:*

$$A \sqsubseteq_k B \iff B^{-1} \sqsubseteq_k A^{-1} \tag{28}$$

A.2 Matrix inverse reverses k_{BA} in same basis case

Theorem 2. *For two density matrices A and B with the same eigenbasis, k_{BA} is reversed by matrix inverse:*

$$k_{BA}(B^{-1}, A^{-1}) = k_{BA}(A, B) \tag{29}$$

Proof.

$$\begin{aligned}
k_{\mathsf{BA}}(\mathsf{B}^{-1}, \mathsf{A}^{-1}) &= \frac{\sum_i \lambda^i_{\mathsf{A}^{-1}} - \lambda^i_{\mathsf{B}^{-1}}}{\sum_i \left|\lambda^i_{\mathsf{A}^{-1}} - \lambda^i_{\mathsf{B}^{-1}}\right|} && (30)\\
&= \frac{\sum_i \frac{1}{\lambda^i_{\mathsf{A}}} - \frac{1}{\lambda^i_{\mathsf{B}}}}{\sum_i \left|\frac{1}{\lambda^i_{\mathsf{A}}} - \frac{1}{\lambda^i_{\mathsf{B}}}\right|} && (31)\\
&= \frac{\sum_i \lambda^i_{\mathsf{B}} - \lambda^i_{\mathsf{A}}}{\sum_i \left|\lambda^i_{\mathsf{B}} - \lambda^i_{\mathsf{A}}\right|} && (32)\\
&= k_{\mathsf{BA}}(\mathsf{A}, \mathsf{B}) && (33)
\end{aligned}$$

using Equation 13 from Equation 30 to 31. □

A.3 Composing with $\neg_{sub}$ or $\neg_{inv}$ gives maximally mixed support

Theorem 3. *When composing a density matrix X with $\neg_{supp}\mathsf{X}$ via* spider, fuzz, *or* phaser, *the resulting density matrix has the desired property of being a maximally mixed state on the support with zeroes on the kernel.*

Proof. $\neg_{supp}\mathsf{X}$ and X have the same eigenbasis. From Equation 13, all nonzero eigenvalues of $\neg_{supp}\mathsf{X}$ are multiplicative inverses of the corresponding eigenvalue of X.
We use definitions of *spider*, *fuzz*, and *phaser* from Equations 1, 2, and 3. The summation indices are over eigenvectors with nonzero eigenvalue.

$$\begin{aligned}
&\textit{spider}(\mathsf{X}, \neg_{supp}\mathsf{X}) && (34)\\
&= U_s(\mathsf{X} \otimes \neg_{supp}\mathsf{X})U_s^\dagger && (35)\\
&= \Big(\sum_i |i\rangle\langle ii|\Big)(\mathsf{X} \otimes \neg_{supp}\mathsf{X})\Big(\sum_j |jj\rangle\langle j|\Big) && (36)\\
&= \sum_i |i\rangle\langle ii| \left(\left(\lambda\, |i\rangle\langle i|\right) \otimes \left(\frac{1}{\lambda_i}\, |i\rangle\langle i|\right)\right) |ii\rangle\langle i| && (37)\\
&= \sum_i |i\rangle\langle i| && (38)\\
&= \mathbb{I}_{supp} && (39)
\end{aligned}$$

$$\begin{aligned}
\textit{fuzz}(\mathsf{X}, \neg_{supp}\mathsf{X}) &= \sum_i x_i P_i \circ \mathsf{X} \circ P_i && (40)\\
&= \sum_i \frac{1}{\lambda_i} P_i \Big(\sum_j \lambda_i P_i\Big) P_i && (41)\\
&= \sum_i P_i && (42)\\
&= \mathbb{I}_{supp} && (43)
\end{aligned}$$

$$\begin{aligned}
&\textit{phaser}(\mathsf{X}, \neg_{supp}\mathsf{X}) && (44)\\
&= \Big(\sum_i x_i P_i\Big) \circ \mathsf{X} \circ \Big(\sum_i x_i P_i\Big) && (45)\\
&= \Big(\sum_i \lambda_i{}^{-\frac{1}{2}} P_i\Big)\Big(\sum_j \lambda_j P_j\Big)\Big(\sum_k \lambda_k{}^{-\frac{1}{2}} P_k\Big) && (46)\\
&= \sum_i P_i && (47)\\
&= \mathbb{I}_{supp} && (48)
\end{aligned}$$

□

Corollary 2. *When composing a density matrix* X *with* $\neg_{inv} X$ *via* spider, fuzz, *or* phaser, *the resulting density matrix has the desired property of being a maximally mixed state on the support with zeroes on the kernel.*