Entailment Relations on Distributions

John van de Wetering
Radboud University
Nijmegen, Netherlands
Department of Computer Science
Oxford University
Oxford, United Kingdom
j.wetering@student.ru.nl

In this paper we give an overview of partial orders on the space of probability distributions that carry a notion of information content and serve as a generalisation of the Bayesian order given in [8]. We investigate what constraints are necessary in order to get a unique notion of information content. These partial orders can be used to give an ordering on words in vector space models of natural language meaning relating to the contexts in which words are used, which is useful for a notion of entailment and word disambiguation. The construction used also points towards a way to create orderings on the space of density operators which allow a more fine-grained study of entailment. The partial orders in this paper are directed complete and form domains in the sense of domain theory.

1 Introduction

Distributional models of natural language form a popular way to study language in the context of automated natural language processing. These models rely on the Distributional Hypothesis: words that occur in similar contexts have similar meanings.

The categorical compositional distributional model of natural language meaning developed by Coecke, Sadrzadeh and Clark [6, 5] also gives a way of composing distributions. It has been corroborated empirically that on some tests this model performs better then the state of the art [9].

Recently this model has been expanded to take into account lexical ambiguity, notions of homonymy and polysemy [12] and entailment at the word and sentence level by passing from a vector space model to a density matrix model [1, 2, 3]. In these papers various relations between the density matrices were explored to get a definition of entailment on the distributional level, such as the fidelity and the relative entropy. In [3] a modified Löwner ordering was used to get a notion of graded entailment. In [10] they constructed a nonsymmetric similarity measure based on a modified measure of feature inclusion.

In any of these cases the goal is to get a relation between pairs of words or sentences that captures the idea of information content. We say that the word *dog* entails the word *animal* because in most contexts where the word *dog* is used, we could use the more general (less specific, less informative) word *animal*.

The same is true for word disambiguation. Consider the word *bank*. It might mean *river bank* or *investment bank*. Without any further context we don't know which one is meant. The word *bank* offers less information than either of these more specific words. We can consider it to be in a mixed state of these pure meanings, which collapses to a pure state when given the right context.

It therefore makes sense that to get a notion of disambiguation or lexical entailment we should be looking for a relation that captures the idea of information content. The obvious properties that we would require of such a relation are those of a partial order: reflexive, transitive and antisymmetric.

Kartsaklis, D., Lewis, M. and Rimell, L. (Eds):
2016 Workshop on Semantic Spaces at the Intersection of
NLP, Physics and Cognitive Science (SLPCS'16)
EPTCS 221, 2016, pp. 58–66, doi:10.4204/EPTCS.221.7

The way word vectors are usually constructed is by counting the coocurrence with some set of basis words. The components can then be interpreted as probabilities of a word occuring at the same time as this specific basis word. So in fact, the word vector can be seen as a probability distribution.

If a word is instead represented by a density matrix then when it is diagonalised we have a probability distribution on the diagonal. This means that a relation capturing the notion of information content should at least be a partial order on the space of probability distributions.

There is a well known partial order on the space of positive semi-definite matrices called the Löwner order, but on the space of density operators no two different density operators are comparable (if we have $x \sqsubseteq y$ then we must have $x = y$). This is a direct effect of the normalisation of the trace of the operators. A modification to the Löwner order was made in [3] in order to get a notion of graded entailment. The resulting structure was no longer a partial order, since the modification removed transitivity and replaced it with a weaker condition. In this paper we will show two different modifications to the Löwner order that do result in proper partial orders.

An example of a nontrivial partial order on the space of probability distributions that has suitable information-like properties is the Bayesian order outlined in [7, 8]. This is in fact the only example the author could find in literature. The Bayesian order served as the inspiration for this paper and the results outlined here can be seen as generalisations of the results related to the Bayesian order.

In this paper we will explore what conditions we need in order for the resulting partial order to represent information content. We will also look at what kind of conditions we need in addition to get a unique notion of information content. Since there has been surprisingly little work in the area of partial orders representing information we will focus on partial orders on probability distributions instead of on the bigger space of density operators. We will also just be looking at entailment on the word level and leave compositionality for further research.

Note also that the results in this paper might prove useful in resource theory and quantum information theory as density matrices are quantum states and probability distributions are classical states. The partial orders studied in this paper turn out to be domains: directed complete partial orders which are exact.

2 Background

We begin by stating the definition of a partial order.

Definition 1. A *partial order* on a space S is a binary relation $\sqsubseteq$ which is

1. Reflexive: $\forall x \in S : x \sqsubseteq x$.
2. Transitive: $\forall x,y,z \in S : x \sqsubseteq y$ and $y \sqsubseteq z \implies x \sqsubseteq z$.
3. Antisymmetric: $\forall x,y \in S : x \sqsubseteq y$ and $y \sqsubseteq x \implies x = y$.

We can restrict a partial order on the density matrices to the diagonal density matrices. This is equivalent to the space of finite probability distributions $\Delta^n = \{(x_1,\ldots,x_n); x_i \geq 0, \sum_i x_i = 1\}$, which can be interpreted geometrically as the $(n-1)$-simplex.

We can then wonder when this procedure can be reversed: which partial orders on the diagonal density matrices extend to a partial order on the entire space of density matrices? The naive approach is to define $\rho \sqsubseteq' \pi$ iff $\mathrm{Diag}(\rho) \sqsubseteq \mathrm{Diag}(\pi)$, where $\mathrm{Diag}(\rho)$ is the probability distribution of the eigenvalues of ρ. However if we take an arbitrary density matrix ρ the diagonalisation will not be fully determined: we can still freely permute the basis vectors. Reflexivity would then imply that any permutation of basis must be equivalent which would in turn break antisymmetry. We must require that ρ and π be diagonalised simultaneously in order for them to be comparable by a partial order on Δ^n.

If two density matrices can be diagonalised simultaneously then there is still a freedom of permuting the basis vectors, so a necessary condition for $\sqsubseteq$ to be extended to the entire space of density operators is for it to be invariant under basis vector permutation:

Definition 2. Let $\sqsubseteq$ be a partial order on Δ^n. We call it *permutation invariant* if for any permutation $\sigma \in S^n$: $x \sqsubseteq y \implies \sigma(x) \sqsubseteq \sigma(y)$

It can be shown that a permutation invariant partial order extends to a partial order on the density matrices (a density matrix is completely determined by its eigenvalues and an orthonormal basis).

A notion of information content is Shannon entropy. On Δ^n the element with the highest amount of entropy is the uniform distribution $\bot = \frac{1}{n}(1,\ldots,1)$. The elements with the lowest amount of entropy are the pointed distributions that have $x_i = 1$ for some i and the rest equal to zero, also called the 'pure' states. Denote these as $\top_i$. Intuitively $\bot$ is the element with the lowest amount of information, and $\top_i$ are the elements with the most amount of information. We require that our partial order on Δ^n respects this: every distribution contains more information than $\bot$ and every distribution is smaller than at least one maximal element.

Linguistically a word would be represented by the uniform distribution if it occured the same amount of times in any context, but such a word would of course not add any information to the sentence. A candidate for such a word would for instance be '*the*'. Realistically no word will be represented by the uniform distribution, but we would find examples of words that are uniformly distributed on a subset of contexts. Such as the word '*bank*' that we would expect somewhat uniformly in the contexts of finance and rivers. Stating that each word can be compared to some pure state is akin to stating that each word can be resolved to some pure meaning.

In order to restrict ourselves to nontrivial partial orders we will require one further property: that the partial order respects the mixing of information content, defined as such:

Definition 3. We say that a partial order on Δ^n allows *mixing* when we have for any x, y and $t \in [0,1]$:

$$x \leq y \implies x \leq (1-t)x + ty \leq y$$

This states that when an element contains less information than another and this information is comparable, then mixing the information content will give something with an information content in between. Note that the space of probability distributions is convex. This demand makes the partial order respect that convexity in a natural way. We are now ready to give a minimal definition of partial order that represents information content.

Definition 4. A partial order on Δ^n which is permutation invariant, allows mixing and has the uniform distribution as the minimal element and the pointed distributions as the maximal elements is an *information ordering*.

There is a unique partial order satisfying the conditions of Definition 4 on Δ^2 as seen in Figure 1. The pure distributions are at the ends while the uniform distribution is in the middle.

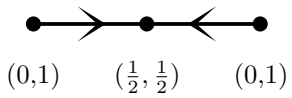


Figure 1: The unique partial order on Δ^2 satisfying Definition 4.

We might hope that these conditions also uniquely determine a partial order for higher values of n, but this is not the case. The inductive procedure in [8] uniquely determines a partial order that does have the right properties, but as we will see we can create other partial orders without using this inductive procedure. The structure that *is* fixed by these conditions is illustrated for $n = 3$ in Figure 2.

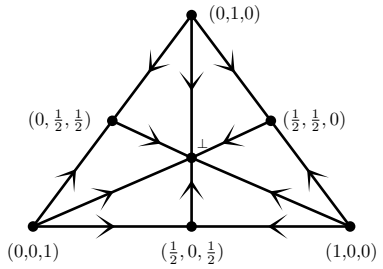


Figure 2: The basic structure of an information ordering on Δ^3.

We see that the space is cut up into natural regions. We will refer to these as sectors, and will come back to those later.

For reference we will state a definition of the Bayesian order here. The other partial orders in this paper will have a similar format.

Definition 5. The Bayesian order $\sqsubseteq_B$ is defined as $x \sqsubseteq_B y$ iff there is a permutation σ such that the coordinates of $\sigma(x)$ and $\sigma(y)$ are both monotonically decreasing and we have
$\sigma(x)_i\sigma(y)_{i+1} \leq \sigma(y)_i\sigma(x)_{i+1}$ for all $1 \leq i \leq n-1$. [8]

The condition that comparable elements must both be able to be permuted in the same way might be seen as odd, but it in fact ensures that the elements are part of the same sector (one of the smaller triangles in Figure 2). As we will see in Section 4, the Bayesian order belongs to a class of partial orders that have this property.

3 Non-Uniqueness of Information Orderings

We will start by showing that the requirements of Definition 4 are not strong enough to give a unique definition of information content. That is: there exist partial orders $\sqsubseteq_1$ and $\sqsubseteq_2$ such that there are points $x \neq y$ with $x \sqsubseteq_1 y$ but $y \sqsubseteq_2 x$.

3.1 Renormalising the Löwner order

As stated in the introduction, the Löwner order given by $x \sqsubseteq y$ iff $y - x \geq 0$ is trivial ($x \sqsubseteq y \implies x = y$) on Δ^n. This is due to the fact that the components of x and y both need to sum up to 1. By renormalising the components so that they no longer sum up to the same value, we are able to create a nontrivial order.

There are at least two natural choices for renormalisation: we can set the largest coordinate equal to 1, or we can set the smallest coordinate equal to 1.

The normalisation to the largest coordinate gives the partial order

$$x \sqsubseteq_L^+ y \iff x^+ y_k \leq y^+ x_k \text{ for all } k.$$

where x^+ is defined as $x^+ = \max\{x_k\}$. This partial order satisfies all the conditions specified in Definition 4, so it is an information ordering.

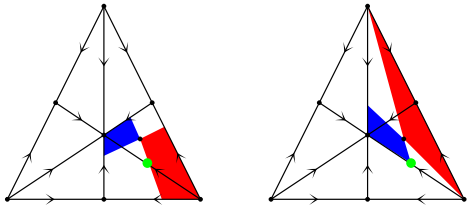
The normalisation to the smallest coordinate is slightly more difficult since the smallest coordinate could be equal to zero. If both elements have the same amount of zeroes we can ignore those and use the smallest nonzero element. If an element y has strictly more zeroes than x we can view y as being blown up to infinity while x stays finite, so we would simply define $x \sqsubseteq y$, as long as their common zeroes are in the same positions.

Keeping this in mind, we can define the second renormalised Löwner order by induction on n as $x \sqsubseteq_{L,n}^- y$ if and only if one of the following holds:

1. There is a k such that $x_k = 0$, $y_k = 0$ and $x \sqsubseteq_{L,(n-1)}^- y$.
2. There is a k such that $y_k = 0$, $x_k = x^-$.
3. For all k: $y_k, x_k \neq 0$ and $x_k y^- \leq y_k x^-$.

Here x^- is defined as the smallest nonzero coordinate.

This is a well-defined partial order and it satisfies all the conditions specified in Definition 4. In Figure 3 you can see that these two renormalisations make a big difference to the resulting partial order. If we take the points $x = \frac{1}{10}(6,2,2)$ and $y = \frac{1}{30}(15,10,5)$, then $x \sqsubseteq_L^+ y$ and $y \sqsubseteq_L^- x$. So there is at least one pair of points where $\sqsubseteq_L^+$ and $\sqsubseteq_L^-$ contradict each other. The conditions specified in Definition 4 are not strong enough to get a unique notion of information content.



(a) $\sqsubseteq_L^+$ (b) $\sqsubseteq_L^-$

Figure 3: Upperset (red) and downset (blue) of the distribution $y = \frac{1}{30}(15,10,5)$ with respect to the renormalised Löwner orders. The point $x = \frac{1}{10}(6,2,2)$ is denoted in green.

3.2 Measurements prevent contradictions

A very useful tool to study the relation between different partial orders are *measurements*, the definition of which we take from [8] and [11].

Definition 6. *A* measurement *is a Scott-continuous strict monotonic map* $\mu : (P, \sqsubseteq_P) \to (S, \sqsubseteq_S)$.

Monotonicity means that when $x \sqsubseteq_P y$ we have $\mu(x) \sqsubseteq_S \mu(y)$ and strictness states that when $x \sqsubseteq_P y$ and $\mu(x) = \mu(y)$ we have $x = y$. Scott-continuity is not important for us, but it is a useful property in relation to proving that a partial order is directed complete. All the strict monotonic maps in this paper are also Scott-continuous.

Define the monotone sector of Δ^n as $\Lambda^n = \{x; x_k \geq x_{k+1}\}$. This corresponds to the lower rightmost triangle in Figure 2. For each $x \in \Delta^n$ there is a unique $y \in \Lambda^n$ such that $y = \sigma(x)$ for some permutation σ. This gives us a natural retraction $r : (\Delta^n, \sqsubseteq) \to (\Lambda^n, \sqsubseteq_{|\Lambda^n})$. r is a measurement for any information ordering $\sqsubseteq$ on Δ^n. This means that if we have a measurement $\mu : \Lambda^n \to S$ this extends to a measurement of Δ^n by composition with r.

The measurements we will be using are of the form $\mu : \Lambda^n \to [0,\infty)^*$.[1] $[0,\infty)^*$ is the positive interval with the reversed order, so monotonicity means that $x \sqsubseteq y$ implies $\mu(x) \geq \mu(y)$.

The order $\sqsubseteq_L^+$ has the measurement $\mu_+(x) = 1 - x^+$. The order $\sqsubseteq_L^-$ has a slightly more complicated measure. Define the zero counting function $Z(x) = \#\{k; x_k = 0\}$. Then when $x \sqsubseteq_L^- y$ we have $Z(y) \geq Z(x)$. If $x \sqsubseteq_L^- y$ and $Z(x) = Z(y)$, then $x^- \geq y^-$, and if additionally $x^- = y^-$ then $x = y$. Putting this together we see that $\mu_-(x) = 2n - 3 - 2Z(x)) + x^-$ is a measurement of $\sqsubseteq_L^-$. We can read this as first counting the amount of zeroes, and then looking at the lowest coordinate. The constant $2n-3$ is added such that $\mu(x) = 0$ iff $x \in \mathrm{Max}(\Delta^n)$.

These two measurements capture different ideas of what we "care" about in our information ordering. Respecting the μ_+ measurement states that the head of a distribution is important, while respecting μ_- means we care about the tail of a distribution.

[1] Any partial order that allows such a measurement is a dcpo [11].

Suppose we have two partial orders $\sqsubseteq_1$ and $\sqsubseteq_2$ that have the same measurement μ. Then if $x \sqsubseteq_1 y$ and $y \sqsubseteq_2 x$ we get $\mu(x) = \mu(y)$ which gives $x = y$. So partial orders with the same measurement can't contradict each other[2]. This gives us a tool for ensuring a class of partial orders won't contradict each other.

4 Restricted Information Orders

We can extend an information order on Δ^n to one on the density operators by allowing comparisons if two density operators can be diagonalised simultaneously. Since we have a measurement from Δ^n to Λ^n we can wonder if we can do the same sort of procedure for transitioning from an information order on Λ^n to one on Δ^n. That is: we allow comparisons when two elements in Δ^n can be brought to Λ^n simultaneously by some permutation $\sigma(x)$. So in that case $\sigma(x), \sigma(y) \in \Lambda^n$ and we proceed with comparing $\sigma(x)$ and $\sigma(y)$ using a partial order on Λ^n. This does however not always result in a valid partial order on Δ^n: Suppose we have $x \sqsubseteq y$ where x is a border element of Λ^n. Then it also lies in a neighbouring sector. Suppose there is an element w in this neighboring sector such that $w \sqsubseteq x$. Then by transitivity $w \sqsubseteq y$. But w and y are in different sectors. So this is a contradiction. A necessary and sufficient condition to prevent this and ensure we can extend a partial order on Λ^n to Δ^n is the following

Definition 7. A partial order $\sqsubseteq$ on Λ^n (or Δ^n) satisfies the *degeneracy condition* when for all $x, y \in \Lambda^n$ (or Δ^n) where $x \sqsubseteq y$ and $y_i = y_j \neq 0$ we have $x_i = x_j \neq 0$.

We call this property the degeneracy condition as it ensures that border elements, elements with a degenerated spectrum, are not above any nondegenerated elements. There is a one-to-one correspondence between information orders satisfying the degeneracy condition on Λ^n and those orders on Δ^n. We will call an information order that satisfies the degeneracy condition a *restricted* information order as comparisons between elements are restricted to within sectors. The renormalised Löwner orders are not restricted information orders, while the Bayesian order *is* a restricted information order.

We are interested in information-like properties of a distribution $x \in \Delta^n$. If we suppose that all these features can be encoded in terms of real numbers, this would give rise to a feature vector $F(x)$ that is an element of $\mathbb{R}^k$ for some k. Comparing the information content of distributions is then translated to comparing the feature vectors of the distributions: $x \sqsubseteq y$ iff $F(x) \leq F(y)$ where $\leq$ is the standard product order on $\mathbb{R}^k$: $v \leq w$ iff $v_i \leq w_i$ for all i. For instance, for $\sqsubseteq_L^+$ the feature vector components are $F(x)_i = x^+/x_i$. For the Bayesian order the feature vector is $F(x)_i = x_i/x_{i+1}$ and for majorization it would be $F(x)_i = \sum_{j=1}^{i} x_j$. We can classify these types of orders.

Theorem 1. Classification of Restricted Information Orderings. All restricted information orderings of the form $x \sqsubseteq y \iff F(x) \leq F(y)$ for some function $F : \Delta^n \to \mathbb{R}^k$ can be written as the join or meet of

[2]Note that two partial orders that do not have the same measurement don't necessarily have to contradict each other.

the set of partial orders $\sqsubseteq_A$ defined as

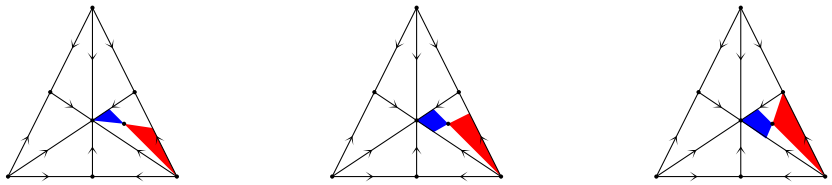
$$\begin{aligned} x \sqsubseteq_A y \iff & f_i(x)g_i(y) \leq f_i(y)g_i(x) \text{ for all } 1 \leq i \leq n-1 \\ & \text{where } f_i(x) = x_i - x_{i+1} \\ & \text{and } g_i(x) = y_{i+1} + \sum_{j=i+2}^{n} A^i_j y_j \text{ where } 1 + \sum_{j=i+2}^{k} A^i_j > 0 \\ & \text{for } 2 \leq i \text{ and } i+1 < k \leq n \\ & \text{and } g_1(x) = y_2 + A^1_0 + \sum_{j=3}^{n} A^1_j y_j \text{ where } 1 + 2A^1_0 > 0 \\ & \text{and } 1 + kA^1_0 + \sum_{j=3}^{k} A^1_j > 0 \text{ for } 2 < k \leq n. \end{aligned}$$

Furthermore, all these partial orders allow μ_- as a measurement, which means they are all dcpo's. The space of these restricted orders is a complete lattice

Note that the feature vectors of these partial orders are $F_i(x) = f_i(x)/g_i(x)$. Using f_i and g_i instead of F_i turns out to be easier because we can deal more naturally with possible zeroes in g_i.

We see that all the parameters A^i_j are bounded from below, but not from above. In general, higher values for the parameters correspond to partial orders that are less strict. The Bayesian order is retrieved when setting all parameters to zero. In general, the restricted orders don't respect the ordering given by Shannon entropy. It can be shown that the subset of restricted orders that allow μ_+ as an additional measurement have Shannon entropy as a measurement as well. All the partial orders seen above also have the property that if $x \sqsubseteq y$ and $x_k = 0$ then $y_k = 0$. Or in other words: the *support* of y is included in x. This ensures that the relative entropy between x and y is finite. These partial orders are therefore somewhat comparable with the entailment relation of [1].

Sharing μ_- as a measurement ensures that these partial orders don't contradict each other. So the degeneracy condition is a sufficient condition to get a unique direction of information content. Because this space of orderings is a complete lattice there is a unique minimal order and a unique maximal order. The difference between these and the Bayesian order is shown in Figure 4.



(a) The minimal order (b) The Bayesian order (c) The maximal order

Figure 4: Upperset (red) and downset (blue) of the same element using the minimal, Bayesian and maximal order on Δ^3.

Note that all the restricted orders and $\sqsubseteq^-_L$ share the measurement μ_-, so they don't contradict each other. It can also be shown that $\sqsubseteq^+_L$ doesn't contradict any restricted order, so both these renormalisations can serve as valid extensions of the restricted orders.

Such an extension is probably necessary. Restricted information orders only allow comparisons within sectors. The amount of sectors in Δ^n is equal to the factorial of n. In an empirical natural language model we would usually have n in the hundreds or thousands, so then there are many more sectors than there are words. This probably means that restricted information orders are too restrictive to be used in practice in the study of natural language as each word will be in its own sector. The renormalised Löwner orders might be better suited to the task.

These partial orders are all constructed by comparing soms feature vectors with each others. This allows for a natural modification to support graded entailment. Suppose we have the partial order $x \sqsubseteq y \iff F(x) \leq F(y)$, then we kan define the p-entailment as $x \sqsubseteq_p y \iff pF(x) \leq F(y)$ for some number $0 \leq p \leq 1$. This is no longer a partial order, but a nonsymmetric entailment measure. This generalises the idea of [3].

5 Information Orders on Density Operators

The central idea behind classifying information orders on probability distributions is that we transition to a feature vector. Let's look at this more closely. We have $\Delta^n \subseteq \mathbb{R}^k$. $\mathbb{R}^k$ has a natural partial order, the product order, which is trivial on Δ^n. By using a feature map F to transform Δ^n to a different subset of $\mathbb{R}^k$, we can make this partial order nontrivial.

The same sort of procedure can be used on the space of density operators $DO(n)$. This space can be seen as a subset of the positive operators $PO(n)$. $PO(n)$ has a natural partial order in the form of the Löwner order, which is trivial on $DO(n)$. We can again consider a "feature map" $F : DO(n) \to PO(n)$ which possibly gives rise to a partial order $\rho \sqsubseteq \pi \iff F(\rho) \sqsubseteq_L F(\pi)$ where $\sqsubseteq_L$ is the Löwner order. For instance, setting $F(\rho) = \rho/\rho^+$ where ρ^+ is the highest eigenvalue of ρ is the natural extension to the density operators of the first renormalised Löwner order described above. In fact, since the Löwner order restricted to diagonal matrices is equal to the product order on $\mathbb{R}^k$, this is a natural generalisation of the construction of information orders on Δ^n.

6 Conclusion and Further Research

We have shown that there is a wide variety of partial orders on the space of probability distributions that satisfy the necessary conditions to capture the notion of information content. With an extra restriction (the degeneracy condition) we can make sure that this notion is unique. Unfortunately in practical linguistic applications this condition might prove to be too strict. The renormalised Löwner orderings are less strict in what they can compare and might prove to be more useful, although empirical research is needed to confirm this. The construction of the restricted information orders also points towards a way to create information orderings on the space of density operators, but studying this in detail is outside of the scope of this paper.

In the pursuit of methods that make comparisons between distributions easier we might look at rescaling distributions to study graded entailment (a generalisation of the approach taken in [3]). Another avenue of attack that might work is using the fact that in a high dimensional space words are probably far apart, so that we can be less picky with the comparisons, and set $x \sqsubseteq y$ whenever some elements within a certain radius of x and y are comparable. This procedure would break antisymmetry when considering the entire space, but not when only comparing words (assuming they are far enough apart). Doing this might allow elements in different sectors to be compared by a restricted information order.

Acknowledgements

The author would like to thank the anonymous reviewers for their valuable feedback. Thanks goes out as well to the authors Master thesis supervisor Bob Coecke of which this publication is part, and to Martha Lewis and Daniel Marsden for valuable comments.

References

[1] Esma Balkir, Dimitri Kartsaklis & Mehrnoosh Sadrzadeh (2015): *Sentence Entailment in Compositional Distributional Semantics*. *arXiv preprint arXiv:1512.04419*. Available at https://arxiv.org/abs/1512.04419.

[2] Esma Balkir, Mehrnoosh Sadrzadeh & Bob Coecke (2016): *Distributional sentence entailment using density matrices*. In: *Topics in Theoretical Computer Science*, Springer, pp. 1–22, doi:10.1007/978-3-319-28678-5_1.

[3] Desislava Bankova, Bob Coecke, Martha Lewis & Daniel Marsden (2016): *Graded Entailment for Compositional Distributional Semantics*. *arXiv preprint arXiv:1601.04908*. Available at https://arxiv.org/abs/1601.04908.

[4] Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do & Chung-chieh Shan (2012): *Entailment Above the Word Level in Distributional Semantics*. In: *Proceedings of the 13th Conference of the European Chapter of the ACL*, EACL '12, pp. 23–32. Available at http://dl.acm.org/citation.cfm?id=2380816.2380822.

[5] Stephen Clark, Bob Coecke & Mehrnoosh Sadrzadeh (2008): *A compositional distributional model of meaning*. In: *Proceedings of the Second Quantum Interaction Symposium (QI-2008)*, pp. 133–140. Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.5435&rep=rep1&type=pdf.

[6] B. Coecke, M. Sadrzadeh & S. Clark (2010): *Mathematical Foundations for a Compositional Distributional Model of Meaning. Lambek Festschrift*. *Linguistic Analysis* 36, pp. 345–384.

[7] Bob Coecke (2003): *Entropic geometry from logic*. *Electronic Notes in Theoretical Computer Science* 83, pp. 39–53, doi:10.1016/S1571-0661(03)50003-2.

[8] Bob Coecke & Keye Martin (2011): *New Structures for Physics*, chapter A Partial Order on Classical and Quantum States, pp. 593–683. Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-642-12821-9_10.

[9] Dimitri Kartsaklis, Mehrnoosh Sadrzadeh & Stephen Pulman (2012): *A Unified Sentence Space for Categorical Distributional-Compositional Semantics: Theory and Experiments*. In: *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Posters, 8-15 December 2012, Mumbai, India*, pp. 549–558. Available at http://aclweb.org/anthology/C/C12/C12-2054.pdf.

[10] Lili Kotlerman, Ido Dagan, Idan Szpektor & Maayan Zhitomirsky-Geffet (2010): *Directional distributional similarity for lexical inference*. *Natural Language Engineering* 16(04), pp. 359–389, doi:10.1017/S1351324910000124.

[11] Keye Martin (2000): *A foundation for computation*. *Tulane University, New Orleans, LA*. Available at http://www.nearmidnight.com/thesis.pdf.

[12] Robin Piedeleu, Dimitri Kartsaklis, Bob Coecke & Mehrnoosh Sadrzadeh (2015): *Open System Categorical Quantum Semantics in Natural Language Processing*. In: *Proceedings of the 6th Conference on Algebra and Coalgebra in Computer Science (CALCO)*, Nijmegen, Netherlands, doi:10.4230/LIPIcs.CALCO.2015.270.