A Framework for Distributional Formal Semantics

Noortje J. Venhuizen[1(✉)], Petra Hendriks[2], Matthew W. Crocker[1], and Harm Brouwer[1]

[1] Saarland University, Saarbrücken, Germany
{noortjev,crocker,brouwer}@coli.uni-saarland.de
[2] University of Groningen, Groningen, The Netherlands
p.hendriks@rug.nl

Abstract. Formal semantics and distributional semantics offer complementary strengths in capturing the meaning of natural language. As such, a considerable amount of research has sought to unify them, either by augmenting formal semantic systems with a distributional component, or by defining a formal system on top of distributed representations. Arriving at such a unified framework has, however, proven extremely challenging. One reason for this is that formal and distributional semantics operate on a fundamentally different 'representational currency': formal semantics defines meaning in terms of models of the world, whereas distributional semantics defines meaning in terms of linguistic co-occurrence. Here, we pursue an alternative approach by deriving a vector space model that defines meaning in a distributed manner relative to formal models of the world. We will show that the resulting *Distributional Formal Semantics* offers probabilistic distributed representations that are also inherently compositional, and that naturally capture quantification and entailment. We moreover show that, when used as part of a neural network model, these representations allow for capturing incremental meaning construction and probabilistic inferencing. This framework thus lays the groundwork for an integrated distributional and formal approach to meaning.

Keywords: Distributionality · Compositionality · Probability · Inference · Incrementality

1 Introduction

In traditional formal semantics, the meaning of a logical expression is typically evaluated in terms of the truth conditions with respect to a formal model M, in which the basic meaning-carrying units (i.e., the basic expressions that are assigned a truth value) are propositions [14]. The meaning of a linguistic expression, then, is defined in terms of the truth conditions its logical translation poses upon a formal model. Critically, these truth conditions define meaning in a segregated manner; distinct propositions obtain separate sets of truth conditions.

R. Iemhoff et al. (Eds.): WoLLIC 2019, LNCS 11541, pp. 633–646, 2019.
https://doi.org/10.1007/978-3-662-59533-6_39

As a result, the relation between individual meanings is not inherently part of their truth-conditional interpretation, but rather follows indirectly from models satisfying these conditions. The core strength of the distributional semantics approach, by contrast, is that (word) meanings are defined in relation to each other, thus directly capturing semantic similarity [18]. It has, however, proven extremely difficult to incorporate well-known features from formal semantics (e.g., compositionality, entailment, etc.) into a distributional semantics framework [3] (but cf. [1,2,7,16,20]).

Here, we take the inverse approach: We introduce distributionality into a formal semantic system, resulting in a framework for Distributional Formal Semantics (DFS). This framework is based on the cognitively inspired meaning representations developed by Golden and Rumelhart [15] and adapted by Frank et al. [12]. In DFS, insights from formal and distributional semantics are combined by defining meaning distributionally over a set of logical models: individual models are treated as observations, or cues, for determining the truth conditions of logical expressions—analogous to how individual linguistic contexts are cues for determining the meaning of words in distributional semantics. Based on a set of logical models $\mathcal{M}$ that together reflect the state of the world both truth-conditionally and probabilistically (i.e., reflecting the probabilistic structure of the world), and a set of propositions $\mathcal{P}$, we can define a vector space for DFS: $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$. The meaning of a proposition is defined as a vector in $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$, which reflects its truth or falsehood relative to each of the models in $\mathcal{M}$. The resulting meaning vector captures the *probabilistic* truth conditions of individual propositions indirectly by identifying the models that satisfy the proposition. Critically, the distributional meaning of individual propositions is defined in relation to all other propositions; propositions that have related meanings will be true in many of the same models, and hence have similar meaning vectors. In other words, the meaning of a proposition is defined in terms of the propositions that it co-occurs with—or, to paraphrase the distributional hypothesis formulated by Firth [10]: "You shall know a *proposition* by the company it keeps".

In what follows, we will show how a well-defined vector space $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ can be derived from a high-level description of the structure of the world, how the resulting meaning space offers distributed representations that are probabilistic and inferential, and how it captures basic concepts from formal semantics, such as compositionality, quantification and entailment. As a proof-of-concept, we present a computational (Prolog) implementation of the DFS framework.[1] Finally, we will show how the DFS representations can be employed in a neural network model for incremental meaning construction. Crucially, we will show how this approach to incremental meaning construction allows for the representation of sub-propositional meaning by exploiting the continuous nature of the meaning space.

[1] DFS-TOOLS is publicly available at http://github.com/hbrouwer/dfs-tools under the Apache License, version 2.0.

2 A Framework for Distributional Formal Semantics

In DFS, the meaning of a proposition $p \in \mathcal{P}$ is defined as a vector $\boldsymbol{v}(p)$ in $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$, such that each unit corresponds to a $M \in \mathcal{M}$, and is assigned a 1 iff M satisfies p, and a 0 otherwise. Consequently, for $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ to be well-defined, the set of models $\mathcal{M}$ that constitutes the meaning space must capture the relevant truth conditions for each proposition $p \in \mathcal{P}$, and conversely, the set of propositions $\mathcal{P}$ must contain all propositions that are captured by each model $M \in \mathcal{M}$. Beyond being well-defined, the meaning space $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ should capture the structure of the world. First of all, the world can enforce hard world knowledge constraints on the co-occurrence of propositions; for instance, certain combinations of propositions may never co-occur, that is, never be simultaneously satisfied within the same model (e.g., a person cannot be at two different places). Secondly, there may be probabilistic constraints on the co-occurrence of propositions; a proposition p may co-occur more frequently with p' than with p'' (for some $p, p', p'' \in \mathcal{P}$), that is, there should be more models $M \in \mathcal{M}$ that satisfy $p \wedge p'$ than models $M' \in \mathcal{M}$ satisfying $p \wedge p''$ (e.g., one prefers reading in bed over reading on the sofa). For $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ to reflect our high-level knowledge about the structure of the world regarding the probabilistic truth-conditions of each proposition $p \in \mathcal{P}$, we thus need its constituent set of models $\mathcal{M}$ to approximate this knowledge. One way of arriving at a satisfactory $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ is to induce this set of models by sampling each model $M \in \mathcal{M}$ from a high-level specification of the structure of the world.

2.1 Sampling $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$

For a given set of propositions $\mathcal{P}$, there are theoretically $2^{\mathcal{P}}$ possible models. Hard constraints in the world rule out any model that satisfies illegal combinations of propositions, while probabilistic constraints require the set of models $\mathcal{M}$ that constitutes $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ to reflect that a proposition p may co-occur more frequently with p' than with p'' (for some $p, p', p'' \in \mathcal{P}$). Hence, the goal is to find a set of models $\mathcal{M}$ such that each $M \in \mathcal{M}$ satisfies all hard constraints, and $\mathcal{M}$ as a whole reflects the probabilistic structure of the world. To this end, we employ an inference-driven, non-deterministic sampling algorithm (inspired by [13]) that stochastically generates models from a set of hard and probabilistic co-occurrence constraints on the propositions $\mathcal{P}$.

As in traditional formal semantics, a model $M \in \mathcal{M}$ is defined as the tuple $\langle U_M, V_M \rangle$, where U_M defines the universe of M, and V_M is the interpretation function that assigns (sets of) entities to the individual constants and properties that constitute $\mathcal{P}$. Given the set of constants $c_1 \ldots c_n$ defined by $\mathcal{P}$, the universe of each $M \in \mathcal{M}$ is defined as $U_M = \{e_1 \ldots e_n\}$, and the interpretation function is initialized to map each constant onto a unique entity: $V_M(c_i) = e_i$. The next step is to stochastically define an interpretation for all propositions in $\mathcal{P}$, while taking into account hard and probabilistic constraints on world structure. To this end, we start out with the initialized interpretation function, which will be incrementally expanded with the interpretation of individual propositions.

We call this interpretation function the Light World[2] (LV_M), to which will be assigned all propositions that are satisfied in M. To facilitate the incremental, inference-driven construction of M, we will in parallel construct a Dark World interpretation function (DV_M), to which will be assigned all propositions that are *not* satisfied in M. Finally, we assume hard constraints to be represented by a set of well-formed formulas $\mathcal{C}$, while probabilistic constraints are represented by a function $Pr(\phi)$ that assigns a probability to a property ϕ. A model M is then sampled by iterating the following steps:

1. Given the constants $c_1 \dots c_n$ defined by $\mathcal{P}$, let $U_M = \{e_1 \dots e_n\}$, $LV_M(c_i) = e_i$ and $DV_M(c_i) = e_i$.
2. Randomly select a proposition $\phi = P(t_1, \dots, t_n)$ from $\mathcal{P}$ that is not yet assigned in LV_M or DV_M.
3. Let LV'_M be the function that extends LV_M with the interpretation of ϕ, such that $\langle t_1, \dots, t_n \rangle \in LV'_M(P)$.
4. Let Light World Consistency $LWC = \top$ iff each constraint in $\mathcal{C}$ is either satisfied by $\langle U_M, LV'_M \rangle$, or if its complement[3] is not satisfied by $\langle U_M, DV_M \rangle$.
5. Let DV'_M be the function that extends DV_M with the interpretation of ϕ , such that $\langle t_1, \dots, t_n \rangle \in DV'_M(P)$.
6. Let Dark World Consistency $DWC = \top$ iff each constraint in $\mathcal{C}$ is either satisfied by $\langle U_M, LV_M \rangle$, or if its complement is not satisfied by $\langle U_M, DV'_M \rangle$.
7. Provided the outcome of step 4 and step 6:
   - $LWC \wedge DWC$: ϕ can be true in both worlds, let $LV_M = LV'_M$ with probability $Pr(\phi)$ and $DV_M = DV'_M$ with probability $1 - Pr(\phi)$;
   - $LWC \wedge \neg DWC$: ϕ can be inferred to the Light World, let $LV_M = LV'_M$;
   - $\neg LWC \wedge DWC$: ϕ can be inferred to the Dark World, let $DV_M = DV'_M$;
   - $\neg LWC \wedge \neg DWC$: ϕ cannot be inferred to either world, meaning the model thus far is inconsistent, and sampling is restarted from step 1.[4]
8. Repeat from step 2 until each proposition in $\mathcal{P}$ is assigned in LV_M or DV_M.
9. If LV_M satisfies each constraint in $\mathcal{C}$, LV_M will be the interpretation function of the resultant model $M = \langle U_M, LV_M \rangle$.

Repeating this sampling procedure n times will yield a set of models $\mathcal{M}$ with cardinality $|\mathcal{M}| = n$. Crucially, while this procedure only samples one model at a time, the probabilistic assignment of non-inferrable propositions to the Light World in step 7 will assure that each probability $Pr(\phi)$ is approximated by the fraction of models in $\mathcal{M}$ that satisfy ϕ, provided that $\mathcal{M}$ is of sufficient size. An efficient implementation of this sampling algorithm is available as part of DFS-TOOLS (see Footnote 1).

[2] cf. The Legend of Zelda: A Link to the Past (Nintendo, 1992).

[3] While a constraint is a well-formed formula that specifies its truth-conditions relative to the Light World (LV_M), its complement specifies its falsehood-conditions relative to the Dark World (DV_M); e.g., the Light Word constraint $\forall x.sleep(x)$ can be proven to be violated if $\exists x.sleep(x)$ is satisfied in the Dark World. See the appendix for a full set of translation rules.

[4] The sampling of inconsistent models strongly depends on the interdependency of the constraints in $\mathcal{C}$ and can be prevented by defining $\mathcal{C}$ in such a way that all combinations of propositions are explicitly handled.

2.2 Formal Properties of $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$

Compositionality. A well-defined semantic space $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ defines the meaning vectors for a set of individual propositions $\mathcal{P}$ relative to a set of logical models $\mathcal{M}$. Given that the meaning vector $\boldsymbol{v}(p)$ of a proposition $p \in \mathcal{P}$ defines its truth values relative to $\mathcal{M}$, we can define the negation $\neg p$ as the vector that assigns 1 to all $M \in \mathcal{M}$ such that p is not satisfied in M, and 0 otherwise:

$$\boldsymbol{v}_i(\neg p) = 1 \text{ iff } M_i \nvDash p \text{ for } 1 \leq i \leq |\mathcal{M}|$$

The meaning of the conjunction $p \wedge q$, given $p, q \in \mathcal{P}$, is defined as the vector $\boldsymbol{v}(p \wedge q)$ that assigns 1 to all $M \in \mathcal{M}$ such that M satisfies both p and q, and 0 otherwise:

$$\boldsymbol{v}_i(p \wedge q) = 1 \text{ iff } M_i \vDash p \text{ and } M_i \vDash q \text{ for } 1 \leq i \leq |\mathcal{M}|$$

Using the negation and conjunction operators, the meaning of any other logical combination of propositions in the semantic space can be defined, thus allowing for meaning vectors representing expressions of arbitrary logical complexity. Critically, these operations also allow for the definition of quantification. Since $\mathcal{P}$ fully describes the set of propositions expressed in $\mathcal{M}$, the (combined) universe of $\mathcal{M}$ ($U_{\mathcal{M}} = \{u_1, \ldots, u_n\}$) directly derives from $\mathcal{P}$. Universal quantification, then, can be formalized as the conjunction over all entities in $U_{\mathcal{M}}$:

$$\boldsymbol{v}_i(\forall x \phi) = 1 \text{ iff } M_i \vDash \phi[x \backslash u_1] \wedge \ldots \wedge \phi[x \backslash u_n] \text{ for } 1 \leq i \leq |\mathcal{M}|$$

Existential quantification, in turn, is formalized as the disjunction over all entities in $U_{\mathcal{M}}$:

$$\boldsymbol{v}_i(\exists x \phi) = 1 \text{ iff } M_i \vDash \phi[x \backslash u_1] \vee \ldots \vee \phi[x \backslash u_n] \text{ for } 1 \leq i \leq |\mathcal{M}|$$

The vectors from $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ are thus fully compositional at the propositional level. Furthermore, in Sect. 3, we will show how sub-propositional meaning can be constructed by incrementally mapping expressions onto vectors in $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$.

Probability. The semantic space $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ is inherently probabilistic, as the meaning vectors for individual propositions in $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ inherently encode their probability. Given a set of models $\mathcal{M}$ that reflects the probabilistic nature of the world, the probability of p can be defined by the number of models that satisfy p, divided by the total number of models:

$$P(p) = |\{M \in \mathcal{M} \mid M \vDash p\}| / |\mathcal{M}|$$

Thus, propositions that are true in a large set of models will obtain a high probability. Given the notion of compositionality discussed above, the probability of $a \wedge b$ can be defined as the probability of the conjunctive vector $\boldsymbol{v}(a \wedge b)$, where a and b may be atomic propositions in $\mathcal{P}$ or any arbitrarily complex combination thereof. Finally, the conditional probability of b given a is defined as:

$$P(b|a) = P(a \wedge b) / P(a)$$

Inference. As described above, the meaning of individual propositions in $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ is defined in terms of their co-occurrence with other propositions. As a result, the vector representations inherently encode how propositions, and logical combinations thereof, are logically related to each other. Entailment, for instance, is reflected in $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ by means of vectors with overlapping truth values; (complex) proposition a entails b ($a \vDash b$) iff b is satisfied by all models that satisfy a. Based on the definition of conditional probability described above, we can moreover formalize probabilistic inference. Intuitively, a high conditional probability of b given a would indicate that b can be inferred from a, since b is satisfied by a large number of models that satisfy a. However, this conditional probability alone is insufficient, as inference requires quantifying the degree to which a increases (or decreases) the probability of b above and beyond its prior probability $P(b)$. We therefore adopt a score for logical inference that factors out this prior probability [11]:

$$inf(b, a) = \begin{cases} \frac{P(b|a) - P(b)}{1 - P(b)} & \text{if } P(b|a) > P(b) \\ \frac{P(b|a) - P(b)}{P(b)} & \text{otherwise} \end{cases}$$

This score yields a value ranging from $+1$ to -1, where $+1$ indicates that (complex) proposition b is perfectly inferred from a (i.e., a entails b; $a \vDash b$), whereas a value of -1 indicates that the negation of b is perfectly inferred from a ($a \vDash \neg b$). Any inference score in between these extremes reflects probabilistic inference in either direction. In what follows, we will employ this notion of inference in a neural network model of incremental meaning construction.

3 Incremental Meaning Construction

The meaning space $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ defines meaning vectors for all propositions in $\mathcal{P}$, and using the compositional operations described above, vectors can be derived for complex logical combinations of propositions. The meaning space also naturally captures sub-propositional meaning. That is, while vectors representing propositional meaning are binary—reflecting truth- and falsehood within models in $\mathcal{M}$—the meaning space itself is continuous, which means that it also captures meanings that are not directly expressible as (combinations of) propositions. We can exploit this continuous nature of $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ to model the word-by-word, context-dependent construction of (sentence-final) propositional meaning. That is, the meaning of a sub-propositional expression is a real-valued vector that defines a point in the vector space, which is positioned in between those points that instantiate the propositional meanings that the expression pertains to. In contrast to traditional semantic approaches, the DFS approach does not define an operation that simply combines the sub-propositional meanings of two subsequent expressions. Rather, sequences of words $w_1 \ldots w_n$ define a trajectory $\langle \boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \rangle$ through $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$, where each $\boldsymbol{v}_i$ represents the (sub-propositional) meaning induced by the sequence of words $w_1 \ldots w_i$; that is, each word w_i induces a meaning in the context of the meaning assigned to its preceding words $w_1 \ldots w_{i-1}$. Sub-propositional meaning thus critically derives from the

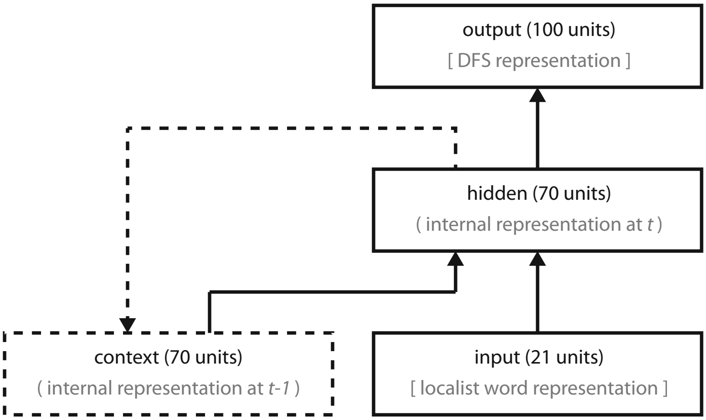

Fig. 1. Simple Recurrent neural Network. Boxes represent groups of artificial neurons, and solid arrows between boxes represent full projections between the neurons in a projecting and a receiving group. The dashed lines indicate that the CONTEXT layer receives a copy of the activation pattern at the HIDDEN layer at the previous time-step. See text for details.

incremental, context-dependent mapping from word sequences onto (complex) propositional meanings. One piece of machinery that is particularly good at approximating such an incremental, context-dependent mapping is the Simple Recurrent neural Network (SRN) [8]. Below, we describe an SRN for incremental meaning construction (cf. [22]) and show how it navigates the meaning space on a word-by-word basis, allowing for incremental (sub-propositional) meaning construction and inferencing.

3.1 Model Specification

We employ an SRN consisting of three groups of artificial logistic dot-product neurons: an INPUT layer (21 units), HIDDEN layer (70), and OUTPUT layer (100) (see Fig. 1). Time in the model is discrete, and at each processing time-step t, activation flows from the INPUT through the HIDDEN layer to the OUTPUT layer. In addition to the activation pattern at the INPUT layer, the HIDDEN layer also receives its own activation pattern at time-step $t-1$ as input (effectuated through an additional CONTEXT layer, which receives a copy of the activation pattern at the HIDDEN layer prior to feedforward propagation). The HIDDEN and the OUTPUT layers both receive input from a bias unit (omitted in Fig. 1). We trained the model using bounded gradient descent [19] to map sequences of localist word representations constituting the words of a sentence, onto a meaning vector from $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ representing the meaning of that sentence.

The sentences on which the model is trained describe situations in a confined world. This world is defined in terms of two persons ($p \in \{john, ellen\}$), two places ($x \in \{restaurant, bar\}$), and two types of food ($f \in \{pizza, fries\}$) and drinks ($d \in \{wine, beer\}$), which can be combined using the following 7 predicates: *enter(p,x)*, *ask_menu(p)*, *order(p,f/d)*, *eat(p,f)*, *drink(p,d)*, *pay(p)* and *leave(p)*. The resulting set of propositions $\mathcal{P}$ ($|\mathcal{P}| = 26$) fully describes the world. A meaning space was constructed from these atomic propositions by sampling a set of 10 K models $\mathcal{M}$ (using the sampling algorithm described in Sect. 2.1), while taking into account world knowledge in terms of hard and probabilistic constraints on proposition co-occurrence; for instance, a person can only enter a single place (hard), and *john* prefers to drink *beer* over *wine* (probabilistic). In order to employ meaning vectors derived from this meaning space in the SRN, we algorithmically selected a subset $\mathcal{M}'$ consisting of 100 models from $\mathcal{M}$, such that $\mathcal{M}'$ adequately reflected the structure of the world (using the algorithm described in [22]). Situations in the world were described using sentences from a language $\mathcal{L}$ consisting of 21 words. The grammar of $\mathcal{L}$ generates a total of 124 sentences, consisting of simple (NP VP) and coordinated (NP VP and VP) sentences. The sentence-initial NPs may be *john*, *ellen*, *someone*, or *everyone*, and the VPs map onto the aforementioned propositions. The corresponding meaning vectors for the sentences in $\mathcal{L}$ were derived using the compositional operations described in Sect. 2.2 (where *someone* and *everyone* correspond to existential and universal quantification, respectively). The model was trained on the full set of sentences generated by $\mathcal{L}$, without any frequency differences between sentences.[5]

Prior to training, the model's weights were randomly initialized using a range of $(-.5, +.5)$. Each training item consisted of a sentence (a sequence of words represented by localist representations) and a meaning vector representing the sentence-final meaning. For each training item, error was backpropagated after each word, using a zero error radius of 0.05, meaning that no error was backpropagated if the error on a unit fell within this radius. Training items were presented in permuted order, and weight deltas were accumulated over epochs consisting of all training items. At the end of each epoch, weights were updated using a learning rate coefficient of 0.1 and a momentum coefficient of 0.9. Training lasted for 5000 epochs, after which the mean squared error was 0.69. The overall performance of the model was assessed by calculating the cosine similarity between each sentence-final output vector and each target vector for all sentences in the training data. All output vectors had the highest cosine similarity to their own target (mean = .99; sd = .02), indicating that the model successfully learned to map sentences onto their corresponding semantics. We moreover computed how well the intended target could be inferred from the output of the model: $\mathit{inf}(\boldsymbol{v}_{target}, \boldsymbol{v}_{output})$.[6] The average inference score over the entire training set was 0.88, which means that after processing a sentence, the model almost perfectly infers the intended meaning of the sentence.

---

[5] The specification of the world described here, including the definition of the language $\mathcal{L}$, is available as part of DFS-TOOLS (see Footnote 1).

[6] For real-valued vectors, we can calculate the probability of vector $\boldsymbol{v}(a)$ as follows: $P(a) = \sum_i \boldsymbol{v}_i(a)/|\mathcal{M}|$.

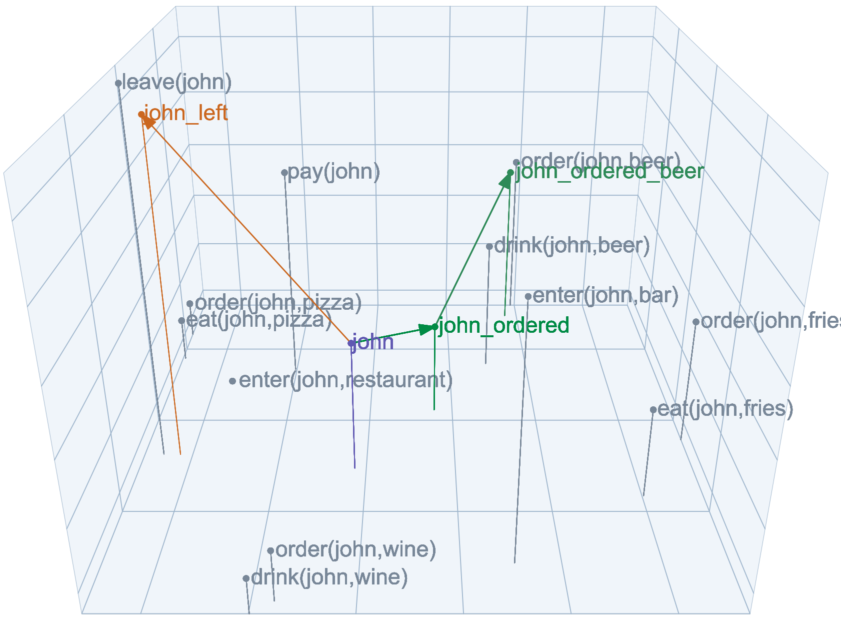

Fig. 2. Visualization of the meaning space into three dimensions (using multidimensional scaling; MDS) for a subset of the atomic propositions (those pertaining to *john*). Grey points represent propositional meaning vectors. Coloured points and arrows show the word-by-word navigational trajectory of the model for the sentences "*john ordered beer*" and "*john left*". See also Footnote 7. (Colour figure online)

3.2 Incremental Inferencing in DFS

On the basis of its linguistic input, the model incrementally constructs a meaning vector at its OUTPUT layer that captures sentence meaning; that is, the model effectively navigates the meaning space $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ on a word-by-word basis. Figure 2 provides a visualization of this navigation process. This figure is a three-dimensional representation of the 100-dimensional meaning space (for a subset of the atomic propositions), derived using multidimensional scaling (MDS). The grey points in this space correspond to propositional meaning vectors. As this figure illustrates, meaning in $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ is defined in terms of co-occurrence; propositions that co-occur frequently in $\mathcal{M}$ (e.g., *order(john,wine)* and *drink(john,wine)*) are positioned close to each other in space.[7] The coloured points show the model's word-by-word output for the sentences "*john ordered beer*" and "*john left*". The navigational trajectory (indicated by the arrows) illustrates how the model assigns intermediate points in meaning space to sub-propositional expressions, and instantiates propositional meanings at

[7] Multidimensional scaling from 100 into 3 dimensions necessarily results in a significant loss of information. Therefore, distances between points in the meaning space shown in Fig. 2 should be interpreted with care.

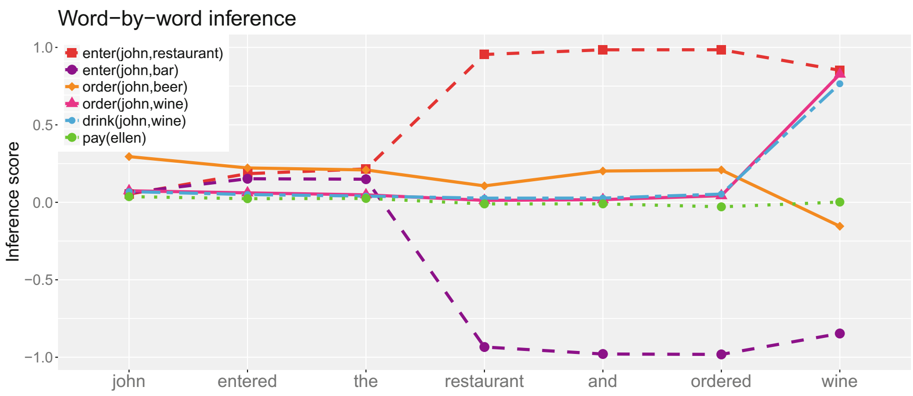

Fig. 3. Word-by-word inference scores of selected propositions for the sentence "*John entered the restaurant and ordered wine*" with the semantics: $enter(john, restaurant) \wedge order(john, wine)$. At a given word, a positive inference score for proposition p indicates that p is positively inferred to be the case; a negative inference score indicates that p is inferred not to be the case (see text for details). (Colour figure online)

sentence-final words. For instance, at the word "*john*", the model navigates to a point in meaning space that is in between the meanings of the propositions pertaining to *john*. The prior probability of propositions in $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ ("world knowledge"), as well as the sentences on which the model was trained ("linguistic experience") together determine the model's trajectory through meaning space. For instance, while the model was exposed to the sentences "*john ordered beer*" and "john ordered wine" equally often, the vector for the expression "*john ordered*" is closer to *order(john,beer)* than *order(john,wine)*, because the former is more probable in the model's knowledge of the world (see [22] for an elaborate investigation of the influence of world knowledge and linguistic experience on meaning space navigation).

Using the inference score described in Sect. 2.2, we can moreover study what the model 'understands' at each word of a sentence (i.e., $inf(b, a)$, where b is the vector of a proposition of interest, and a the output vector of the SRN). Figure 3 shows the word-by-word inference scores for the sentence "*john entered the restaurant and ordered wine*" with respect to 6 propositions. First of all, this figure shows that by the end of the sentence, the model has understood its meaning: the inference scores of *enter(john,restaurant)* and *order(john,wine)* are both ≈ 1 at the sentence-final word. What is more, it does so on an incremental basis: at the word "restaurant", the model commits to the inference *enter(john,restaurant)*, which rules out *enter(john,bar)* since these do not co-occur in the world ($P(enter(john,restaurant) \wedge enter(john,bar)) = 0$). At the word "ordered", the model finds itself in state that is closer to the inference that *order(john,beer)* than *order(john,wine)*, as John prefers beer over wine ($P(order(john,beer)) = 0.81 > P(order(john,wine)) = 0.34$). However, at the word "*wine*" this inference is reversed, and the model understands that *order(john,wine)* is the case, and that *order(john,beer)* cannot be inferred. In addition, the word "*wine*" also leads the model to infer *drink(john,wine)*, even though this proposition is not explicitly part of the semantics of the sentence. This happens because the world stipulates that given that John ordered wine, it is likely that he also drank it ($P(\mathit{drink(john,wine)} \mid \mathit{order(john,wine)}) = 0.88$). Finally, no significant inferences are drawn about the unrelated proposition $pay(ellen)$.

4 Discussion

The DFS framework defines the meaning of a proposition p in terms of models that satisfy it and those that do not. Hence, the framework relies on finding a set of models $\mathcal{M}$ that truth-conditionally and probabilistically capture the structure of the world with respect to a set of propositions $\mathcal{P}$. Here, we focused on how this space $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ can be induced from a high-level description of the structure of the world. We would like to emphasize, however, that none of the described formal properties of the meaning space hinges upon this sampling procedure. An alternative approach towards arriving at $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$, for instance, is to induce it empirically from a semantically annotated corpus (e.g., [4]) or from crowd-sourced human data on propositional co-occurrence (e.g., [23]). The only requirements are that the resultant space $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ is well-defined, and that it accurately approximates the structure of the world in terms of hard and probabilistic constraints on propositional co-occurrence.

DFS representations are inherently compositional at the level of propositions in that atomic propositions can be compositionally combined into complex propositions. At the sub-propositional level, however, meaning is constructed by incrementally navigating $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$. Arriving at the meaning of "john ordered" does not simply involve combining the meaning of "john" with the meaning of "ordered", but rather entails the context-dependent integration of the word "ordered" into the meaning representation constructed after processing "john" (cf. [5]). Crucially, this is possible due to the continuous nature of $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$. Hence, in the DFS framework, compositionality at the propositional level and incrementality at the sub-propositional level interact in context-dependent meaning construction.

The relatively simple neural network model presented here served to illustrate the incremental meaning construction procedure. More sophisticated models, however, instantiating earlier formulations of the DFS framework (cf. [12]), have already highlighted various other interesting properties of the approach. For one, while the current model was trained and tested on the same sentence-semantics pairs, other models have shown generalization to unseen sentences and semantics, in both comprehension [11] and production [6]. Crucially, this semantic systematicity derives from the structure of the world as encoded by the meaning space. Moreover, since in a comprehension model—such as the one described here—each word serves as a contextualized cue for meaning space navigation, a relatively simple SRN architecture (as compared to more complex architectures such as Long Short-Term Memory, LSTM, [17] networks), suffices for this systematicity to manifest. Secondly, other models have explored the dynamics of meaning-space navigation using information-theoretic notions such as surprisal and entropy [13,22].

In DFS, there are two levels at which semantic phenomena can be modeled: the level of the meaning space $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$, and the mapping from words onto points within this meaning space. Starting with the meaning space itself, one could explore different schemes for encoding the atomic propositions, for instance to explicitly capture tense and aspect, or Davidsonian event semantics. Moreover, by varying temporally-dependent proposition co-occurrence within and across models, we obtain different encodings of time within $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ (see [22] for a within-model approach). At the level of the mapping between words and points in $\mathcal{S}_{\mathcal{M}\times\mathcal{P}}$ space, in turn, the DFS framework allows for different ways to capture discourse-level phenomena, such as modality, reference, information structure, and implicature. Crucially, the fact that inference directly follows from incremental semantic meaning construction circumvents the need for a separate pragmatic inference mechanism. This thus blurs the strict line between semantics and pragmatics, thereby directly implementing recent theorizing in formal semantics [21].

While the DFS framework combines formal and distributional approaches to meaning, we take the framework to be complementary to lexically-driven distributional semantics (e.g., LSA; [18]). In DFS, the 'representational currency' is *propositions*, whereas in distributional semantics it is *words*. As a result, DFS allows us to model similarity at the propositional level (e.g., $order(john, beer)$ is similar to $drink(john, beer)$ as they co-occur in $\mathcal{M}$), while distributional semantics models lexical similarity ("*beer*" is similar to "*wine*" as they occur in similar linguistic contexts; e.g., [9]). Crucially, the DFS approach and distributional semantics thus capture different notions of semantic similarity: while the latter offers representations that inherently encode feature-based lexical similarity between words, the former provides representations instantiating the truth-conditional similarity between propositions. The complementary nature of these meaning representations is underlined by recent advances in the neurocognition of language, where evidence suggests that lexical retrieval (the mapping of words onto lexical semantics) and semantic integration (the integration of word meaning into the unfolding representation of propositional meaning) are two distinct processes involved in word-by-word sentence processing [5]. Crucially, this perspective on language comprehension suggests that compositionality is only at play at the level of propositions, thus eschewing the need for compositionality at the lexical level.

5 Conclusion

The DFS framework offers a novel approach to distributional semantics, by defining the meaning of propositions distributionally over a set of formal models. As a consequence, the approach inherits the entire apparatus of (first-order) logic that powers formal semantics, while offering contextualized and probabilistic distributed meaning representations similar to distributional semantics. Crucially, the meaning representations differ from those from distributional semantics in that they offer probabilistic information that reflects the state of the world, rather than linguistic co-occurrence, thereby offering a complementary perspective on meaning representation. To illustrate the approach, we have shown how the DFS meaning space can be derived from a high-level specification of the world, and how it naturally captures well-known concepts from formal semantics, such as compositionality and entailment. Moreover, when employed in an incremental model of meaning construction, it naturally captures sub-propositional meaning and inferencing. As such, we believe that the DFS framework—implemented by DFS-TOOLS—offers a powerful synergy between formal and distributional approaches that paves the way towards novel investigations into formal meaning representation and construction.

Appendix

The complement of any well-formed formula is found by recursively applying the following translations, where ϕ' is the complement of ϕ:

$$\begin{array}{lll} \neg\phi \mapsto \neg\phi' & \phi \veebar \psi \mapsto (\phi' \wedge \psi') \vee (\neg\phi' \wedge \neg\psi') & \exists x.\phi \mapsto \forall x.\phi' \\ \phi \wedge \psi \mapsto \phi' \vee \psi' & \phi \rightarrow \psi \mapsto \neg\phi' \vee \psi' & \forall x.\phi \mapsto \exists x.\phi' \\ \phi \vee \psi \mapsto \phi' \wedge \psi' & \phi \leftrightarrow \psi \mapsto (\neg\phi' \wedge \psi') \vee (\phi' \wedge \neg\psi') & p \mapsto p \end{array}$$

References

1. Baroni, M., Bernardi, R., Zamparelli, R.: Frege in space: a program of compositional distributional semantics. Linguist. Issues Lang. Technol. (LiLT) **9**, 241–346 (2014)
2. Baroni, M., Zamparelli, R.: Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 1183–1193. Association for Computational Linguistics (2010)
3. Boleda, G., Herbelot, A.: Formal distributional semantics: introduction to the special issue. Comput. Linguist. **42**(4), 619–635 (2016)
4. Bos, J., Basile, V., Evang, K., Venhuizen, N.J., Bjerva, J.: The Groningen Meaning Bank. In: Ide, N., Pustejovsky, J. (eds.) Handbook of Linguistic Annotation, pp. 463–496. Springer, Dordrecht (2017). https://doi.org/10.1007/978-94-024-0881-2_18
5. Brouwer, H., Crocker, M.W., Venhuizen, N.J., Hoeks, J.C.J.: A neurocomputational model of the N400 and the P600 in language processing. Cogn. Sci. **41**, 1318–1352 (2017). https://doi.org/10.1111/cogs.12461
6. Calvillo, J., Brouwer, H., Crocker, M.W.: Connectionist semantic systematicity in language production. In: Papafragou, A., Grodner, D., Mirman, D., Trueswell, J.C. (eds.) Proceedings of the 38th Annual Conference of the Cognitive Science Society, Austin, TX, pp. 2555–3560 (2016)
7. Coecke, M.S.B., Clark, S.: Mathematical foundations for a compositional distributed model of meaning. In: Lambek Festschrift, Linguistic Analysis, vol. 36 (2010)

8. Elman, J.L.: Finding structure in time. Cogn. Sci. **14**(2), 179–211 (1990)
9. Erk, K.: What do you know about an alligator when you know the company it keeps? Semant. Pragmat. **9**(17), 1–63 (2016). https://doi.org/10.3765/sp.9.17
10. Firth, J.R.: A synopsis of linguistic theory, 1930–1955. In: Studies in linguistic analysis. Philological Society, Oxford (1957)
11. Frank, S.L., Haselager, W.F.G., van Rooij, I.: Connectionist semantic systematicity. Cognition **110**(3), 358–379 (2009)
12. Frank, S.L., Koppen, M., Noordman, L.G.M., Vonk, W.: Modeling knowledge-based inferences in story comprehension. Cogn. Sci. **27**(6), 875–910 (2003)
13. Frank, S.L., Vigliocco, G.: Sentence comprehension as mental simulation: an information-theoretic perspective. Information **2**(4), 672–696 (2011)
14. Frege, G.: Über Sinn und Bedeutung. Zeitschrift für Philosophie und philosophische Kritik **100**, 25–50 (1892)
15. Golden, R.M., Rumelhart, D.E.: A parallel distributed processing model of story comprehension and recall. Discourse Process. **16**(3), 203–237 (1993)
16. Grefenstette, E., Sadrzadeh, M.: Experimental support for a categorical compositional distributional model of meaning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1394–1404. Association for Computational Linguistics (2011)
17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
18. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. **104**(2), 211–240 (1997)
19. Rohde, D.L.T.: A connectionist model of sentence comprehension and production. Ph.D. thesis, Carnegie Mellon University (2002)
20. Socher, R., Huval, B., Manning, C.D., Ng, A.Y.: Semantic compositionality through recursive matrix-vector spaces. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1201–1211. Association for Computational Linguistics (2012)
21. Venhuizen, N.J., Bos, J., Hendriks, P., Brouwer, H.: Discourse semantics with information structure. J. Semant. **35**(1), 127–169 (2018). https://doi.org/10.1093/jos/ffx017
22. Venhuizen, N.J., Crocker, M.W., Brouwer, H.: Expectation-based comprehension: modeling the interaction of world knowledge and linguistic experience. Discourse Process. **56**(3), 229–255 (2019). https://doi.org/10.1080/0163853X.2018.1448677
23. Wanzare, L.D., Zarcone, A., Thater, S., Pinkal, M.: DeScript: a crowdsourced corpus for the acquisition of high-quality script knowledge. In: The International Conference on Language Resources and Evaluation (2016)