Functional Distributional Semantics

Guy Emerson and Ann Copestake
Computer Laboratory
University of Cambridge
{gete2,aac10}@cam.ac.uk

Abstract

Vector space models have become popular in distributional semantics, despite the challenges they face in capturing various semantic phenomena. We propose a novel probabilistic framework which draws on both formal semantics and recent advances in machine learning. In particular, we separate predicates from the entities they refer to, allowing us to perform Bayesian inference based on logical forms. We describe an implementation of this framework using a combination of Restricted Boltzmann Machines and feedforward neural networks. Finally, we demonstrate the feasibility of this approach by training it on a parsed corpus and evaluating it on established similarity datasets.

1 Introduction

Current approaches to distributional semantics generally involve representing words as points in a high-dimensional vector space. However, vectors do not provide 'natural' composition operations that have clear analogues with operations in formal semantics, which makes it challenging to perform inference, or capture various aspects of meaning studied by semanticists. This is true whether the vectors are constructed using a *count* approach (e.g. Turney and Pantel, 2010) or an *embedding* approach (e.g. Mikolov et al., 2013), and indeed Levy and Goldberg (2014b) showed that there are close links between them. Even the tensorial approach described by Coecke et al. (2010) and Baroni et al. (2014), which naturally captures argument structure, does not allow an obvious account of context dependence, or logical inference.

In this paper, we build on insights drawn from formal semantics, and seek to learn representations which have a more natural logical structure, and which can be more easily integrated with other sources of information.

Our contributions in this paper are to introduce a novel framework for distributional semantics, and to describe an implementation and training regime in this framework. We present some initial results to demonstrate that training this model is feasible.

2 Formal Framework of Functional Distributional Semantics

In this section, we describe our framework, explaining the connections to formal semantics, and defining our probabilistic model. We first motivate representing predicates with functions, and then explain how these functions can be incorporated into a representation for a full utterance.

2.1 Semantic Functions

We begin by assuming an extensional model structure, as standard in formal semantics (Kamp and Reyle, 1993; Cann, 1993; Allan, 2001). In the simplest case, a model contains a set of entities, which predicates can be true or false of. Models can be endowed with additional structure, such as for plurals (Link, 2002), although we will not discuss such details here. For now, the important point is that we should separate the representation of a predicate from the representations of the entities it is true of.

We generalise this formalisation of predicates by treating truth values as random variables,[1]

[1]The move to replace absolute truth values with probabilities has parallels in much computational work based on formal logic. For example, Garrette et al. (2011) incorporate distributional information in a Markov Logic Network (Richardson and Domingos, 2006). However, while their approach allows probabilistic inference, they rely on existing distributional vectors, and convert similarity scores to weighted logical formulae. Instead, we aim to learn representations which are directly interpretable within in a probabilistic logic.

Proceedings of the 1st Workshop on Representation Learning for NLP, pages 40–52,
Berlin, Germany, August 11th, 2016.

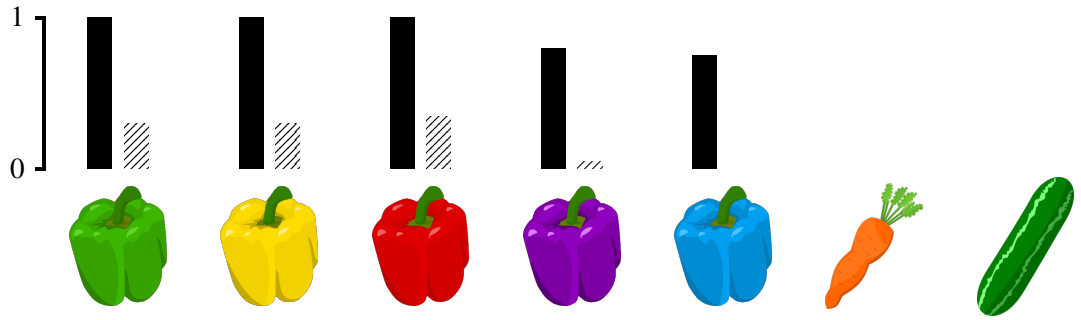

Figure 1: Comparison between a *semantic function* and a *distribution* over a space of entities. The vegetables depicted above (five differently coloured bell peppers, a carrot, and a cucumber) form a discrete semantic space $\mathcal{X}$. We are interested in the truth t of the predicate for *bell pepper* for an entity $x \in \mathcal{X}$. Solid bars: the semantic function $P(t|x)$ represents how much each entity is considered to be a pepper, and is bounded between 0 and 1; it is high for all the peppers, but slightly lower for atypical colours. Shaded bars: the distribution $P(x|t)$ represents our belief about an entity if all we know is that the predicate for *bell pepper* applies to it; the probability mass must sum to 1, so it is split between the peppers, skewed towards typical colours, and excluding colours believed to be impossible.

which enables us to apply Bayesian inference. For any entity, we can ask which predicates are true of it (or 'applicable' to it). More formally, if we take entities to lie in some semantic space $\mathcal{X}$ (whose dimensions may denote different features), then we can take the meaning of a predicate to be a function from $\mathcal{X}$ to values in the interval $[0, 1]$, denoting how likely a speaker is to judge the predicate applicable to the entity. This judgement is variable between speakers (Labov, 1973), and for borderline cases, it is even variable for one speaker at different times (McCloskey and Glucksberg, 1978).

Representing predicates as functions allows us to naturally capture vagueness (a predicate can be equally applicable to multiple points), and using values between 0 and 1 allows us to naturally capture gradedness (a predicate can be more applicable to some points than to others). To use Labov's example, the predicate for *cup* is equally applicable to vessels of different shapes and materials, but becomes steadily less applicable to wider vessels.

We can also view such a function as a classifier – for example, the semantic function for the predicate for *cat* would be a classifier separating cats from non-cats. This ties in with a view of concepts as abilities, as proposed in both philosophy (Dummett, 1978; Kenny, 2010), and cognitive science (Murphy, 2002; Bennett and Hacker, 2008). A similar approach is taken by Larsson (2013), who argues in favour of representing perceptual concepts as classifiers of perceptual input.

Note that these functions do not directly define probability distributions over entities. Rather, they define binary-valued *conditional* distributions, *given an entity*. We can write this as $P(t|x)$, where x is an entity, and t is a stochastic truth value. It is only possible to get a corresponding distribution over entities given a truth value, $P(x|t)$, if we have some background distribution $P(x)$. If we do, we can apply Bayes' Rule to get $P(x|t) \propto P(t|x)P(x)$. In other words, the truth of an expression depends crucially on our knowledge of the situation. This fits neatly within a verificationist view of truth, as proposed by Dummett (1976), who argues that to understand a sentence is to know how we could verify or falsify it.

By using both $P(t|x)$ and $P(x|t)$, we can distinguish between underspecification and uncertainty as two kinds of 'vagueness'. In the first case, we want to state partial information about an entity, but leave other features unspecified; $P(t|x)$ represents which kinds of entity could be described by the predicate, regardless of how likely we think the entities are. In the second case, we have uncertain knowledge about the entity; $P(x|t)$ represents which kinds of entity we think are likely for this predicate, given all our world knowledge.

For example, bell peppers come in many colours, most typically green, yellow, orange or red. As all these colours are typical, the semantic function for the predicate for *bell pepper* would take a high value for each. In contrast, to define a probability distribution over entities, we must split probability mass between different colours,[2]

[2]In fact, colour would be most properly treated as a continuous feature. In this case, $P(x)$ must be a probability density function, not a probability mass function, whose value would further depend on the parametrisation of the space.

and for a large number of colours, we would only have a small probability for each. As purple and blue are atypical colours for a pepper, a speaker might be less willing to label such a vegetable a pepper, but not completely unwilling to do so – this linguistic knowledge belongs to the semantic function for the predicate. In contrast, after observing a large number of peppers, we might conclude that blue peppers do not exist, purple peppers are rare, green peppers common, and red peppers more common still – this world knowledge belongs to the probability distribution over entities. The contrast between these two quantities is depicted in figure 1, for a simple discrete space.

2.2 Incorporation with Dependency Minimal Recursion Semantics

Semantic dependency graphs have become popular in NLP. We use Dependency Minimal Recursion Semantics (DMRS) (Copestake et al., 2005; Copestake, 2009), which represents meaning as a directed acyclic graph: nodes represent predicates/entities (relying on a one-to-one correspondence between them) and links (edges) represent argument structure and scopal constraints. Note that we assume a neo-Davidsonian approach (Davidson, 1967; Parsons, 1990), where events are also treated as entities, which allows a better account of adverbials, among other phenomena.

For example (simplifying a little), to represent *"the dog barked"*, we have three nodes, for the predicates *the*, *dog*, and *bark*, and two links: an ARG1 link from *bark* to *dog*, and a RSTR link from *the* to *dog*. Unlike syntactic dependencies, DMRS abstracts over semantically equivalent expressions, such as *"dogs chase cats"* and *"cats are chased by dogs"*. Furthermore, unlike other types of semantic dependencies, including Abstract Meaning Representations (Banarescu et al., 2012), and Prague Dependencies (Böhmová et al., 2003), DMRS is interconvertible with MRS, which can be given a direct logical interpretation.

We deal here with the extensional fragment of language, and while we can account for different quantifiers in our framework, we do not have space to discuss this here – for the rest of this paper, we neglect quantifiers, and the reader may assume that all variables are existentially quantified.

We can use the structure of a DMRS graph to define a probabilistic graphical model. This gives us a distribution over lexicalisations of the graph – given an abstract graph structure, where links are labelled but nodes are not, we have a process to generate a predicate for each node. Although this process is different for each graph structure, we can share parameters between them (e.g. according to the labels on links). Furthermore, if we have a distribution over graph structures, we can incorporate that in our generative process, to produce a distribution over lexicalised graphs.

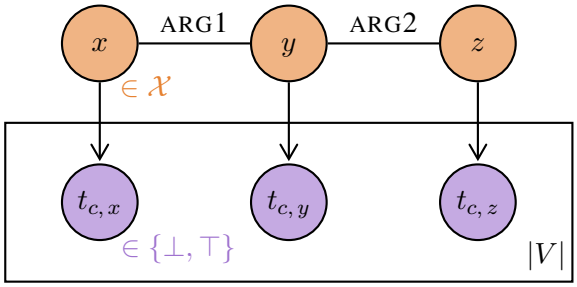

Figure 2: A situation composed of three entities. Top row: the entities x, y, and z lie in a semantic space $\mathcal{X}$, jointly distributed according to DMRS links. Bottom row: each predicate c in the vocabulary V has a stochastic truth value for each entity.

The entity nodes can be viewed as together representing a situation, in the sense of Barwise and Perry (1983). We want to be able to represent the entities without reference to the predicates – intuitively, the world is the same however we choose to describe it. To avoid postulating causal structure amongst the entities (which would be difficult for a large graph), we can model the entity nodes as an undirected graphical model, with edges according to the DMRS links. The edges are undirected in the sense that they don't impose conditional dependencies. However, this is still compatible with having 'directed' semantic dependencies – the probability distributions are not symmetric, which maintains the asymmetry of DMRS links.

Each node takes values in the semantic space $\mathcal{X}$, and the network defines a joint distribution over entities, which represents our knowledge about which situations are likely or unlikely. An example is shown in the top row of figure 2, of an entity y along with its two arguments x and z – these might represent an event, along with the agent and patient involved in the event. The structure of the graph means that we can factorise the joint distribution $P(x, y, z)$ over the entities as being proportional to the product $P(x, y)\, P(y, z)$.

For any entity, we can ask which predicates are true of it. We can therefore introduce a

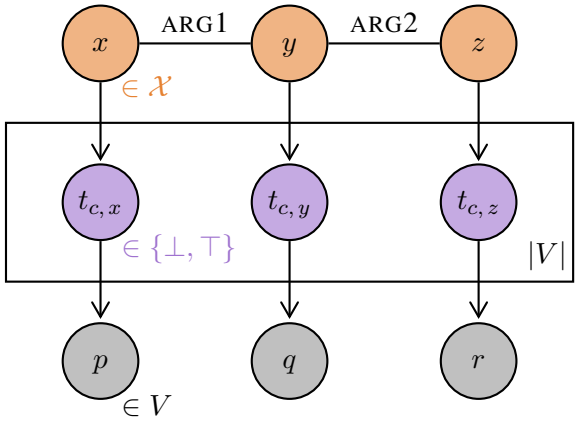

Figure 3: The probabilistic model in figure 2, extended to generate utterances. Each predicate in the bottom row is chosen out of all predicates which are true for the corresponding entity.

node for every predicate in the vocabulary, where the value of the node is either true ($\top$) or false ($\bot$). Each of these predicate nodes has a single directed link from the entity node, with the probability of the node being true being determined by the predicate's semantic function, i.e. $P(t_{c,x} = \top | x) = t_c(x)$. This is shown in the second row of figure 2, where the plate denotes that these nodes are repeated for each predicate c in the vocabulary V. For example, if the situation represented a dog chasing a cat, then nodes like $t_{dog,x}$, $t_{animal,x}$, and $t_{pursue,y}$ would be true (with high probability), while $t_{democracy,x}$ or $t_{dog,z}$ would be false (with high probability).

The probabilistic model described above closely matches traditional model-theoretic semantics. However, while we could stop our semantic description there, we do not generally observe truth-value judgements for all predicates at once;[3] rather, we observe utterances, which have specific predicates. We can therefore define a final node for each entity, which takes values over predicates in the vocabulary, and which is conditionally dependent on the truth values of all predicates. This is shown in the bottom row of figure 3. Including these final nodes means that we can train such a model on observed utterances. The process of choosing a predicate from the true ones may be complex, potentially depending on speaker intention and other pragmatic factors – but in section 3, we will simply choose a true predicate at random (weighted by frequency).

[3]This corresponds to what Copestake and Herbelot (2012) call an *ideal distribution*. If we have access to such information, we only need the two rows given in figure 2.

The separation of entities and predicates allows us to naturally capture context-dependent meanings. Following the terminology of Quine (1960), we can distinguish context-independent *standing* meaning from context-dependent *occasion* meaning. Each predicate type has a corresponding semantic function – this represents its standing meaning. Meanwhile, each predicate token has a corresponding entity, for which there is a posterior distribution over the semantic space, conditioning on the rest of the graph and any pragmatic factors – this represents its occasion meaning.

Unlike previous approaches to context dependence, such as Dinu et al. (2012), Erk and Padó (2008), and Thater et al. (2011), we represent meanings in and out of context by different kinds of object, reflecting a type/token distinction. Even Herbelot (2015), who explicitly contrasts individuals and kinds, embeds both in the same space.

As an example of how this separation of predicates and entities can be helpful, suppose we would like *"dogs chase cats"* and *"cats chase mice"* to be true in a model, but *"dogs chase mice"* and *"cats chase cats"* to be false. In other words, there is a dependence between the verb's arguments. If we represent each predicate by a single vector, it is not clear how to capture this. However, by separating predicates from entities, we can have two different entities which *chase* is true of, where one co-occurs with a dog-entity ARG1 and cat-entity ARG2, while the other co-occurs with a cat-entity ARG1 and a mouse-entity ARG2.

3 Implementation

In the previous section, we described a general framework for probabilistic semantics. Here we give details of one way that such a framework can be implemented for distributional semantics, keeping the architecture as simple as possible.

3.1 Network Architecture

We take the semantic space $\mathcal{X}$ to be a set of binary-valued vectors,[4] $\{0,1\}^N$. A situation s is then composed of entity vectors $x^{(1)}, \cdots, x^{(K)} \in \mathcal{X}$ (where the number of entities K may vary), along with links between the entities. We denote a link from $x^{(n)}$ to $x^{(m)}$ with label l as: $x^{(n)} \xrightarrow{l} x^{(m)}$. We define the background distribution over situations using a Restricted Boltzmann Machine

[4]We use the term *vector* in the computer science sense of a linear array, rather than in the mathematical sense of a point in a vector space.

(RBM) (Smolensky, 1986; Hinton et al., 2006), but rather than having connections between hidden and visible units, we have connections between components of entities, according to the links.

The probability of the network being in the particular configuration s depends on the *energy* of the configuration, $E^b(s)$, as shown in equations (1)-(2). A high energy denotes an unlikely configuration. The energy depends on the edges of the graphical model, plus bias terms, as shown in (3). Note that we follow the Einstein summation convention, where repeated indices indicate summation; although this notation is not typical in NLP, we find it much clearer than matrix-vector notation, particularly for higher-order tensors. Each link label l has a corresponding weight matrix $W^{(l)}$, which determines the strength of association between components of the linked entities. The first term in (3) sums these contributions over all links $x^{(n)} \xrightarrow{l} x^{(m)}$ between entities. We also introduce bias terms, to control how likely an entity vector is, independent of links. The second term in (3) sums the biases over all entities $x^{(n)}$.

$$P(s) = \frac{1}{Z} \exp\left(-E^b(s)\right) \tag{1}$$

$$Z = \sum_{s'} \exp\left(-E^b(s')\right) \tag{2}$$

$$-E^b(s) = \sum_{x^{(n)} \xrightarrow{l} x^{(m)}} W^{(l)}_{ij} x^{(n)}_i x^{(m)}_j - \sum_{x^{(n)}} b_i x^{(n)}_i \tag{3}$$

Furthermore, since sparse representations have been shown to be beneficial in NLP, both for applications and for interpretability of features (Murphy et al., 2012; Faruqui et al., 2015), we can enforce sparsity in these entity vectors by fixing a specific number of units to be active at any time. Swersky et al. (2012) introduce this RBM variant as the Cardinality RBM, and also give an efficient exact sampling procedure using belief propagation. Since we are using sparse representations, we also assume that all link weights are non-negative.

Now that we've defined the background distribution over situations, we turn to the semantic functions t_c, which map entities x to probabilities. We implement these as feedforward networks, as shown in (4)-(5). For simplicity, we do not introduce any hidden layers. Each predicate c has a vector of weights $W'^{(c)}$, which determines the strength of association with each dimension of the semantic space, as well as a bias term $b'^{(c)}$. These together define the energy $E^p(x, c)$ of an entity x with the predicate, which is passed through a sigmoid function to give a value in the range $[0, 1]$.

$$t_c(x) = \sigma(-E^p(x, c)) = \frac{1}{1 + \exp\left(E^p\right)} \tag{4}$$

$$-E^p(x, c) = W'^{(c)}_i x_i - b'^{(c)} \tag{5}$$

Given the semantic functions, choosing a predicate for a entity can be hard-coded, for simplicity. The probability of choosing a predicate c for an entity x is weighted by the predicate's frequency f_c and the value of its semantic function $t_c(x)$ (how true the predicate is of the entity), as shown in (6)-(7). This is a mean field approximation to the stochastic truth values shown in figure 3.

$$P(c|x) = \frac{1}{Z_x} f_c t_c(x) \tag{6}$$

$$Z_x = \sum_{c'} f_{c'} t_{c'}(x) \tag{7}$$

3.2 Learning Algorithm

To train this model, we aim to maximise the likelihood of observing the training data – in Bayesian terminology, this is *maximum a posteriori* estimation. As described in section 2.2, each data point is a lexicalised DMRS graph, while our model defines distributions over lexicalisations of graphs. In other words, we take as given the observed distribution over abstract graph structures (where links are given, but nodes are unlabelled), and try to optimise how the model generates predicates (via the parameters $W^{(l)}_{ij}, b_i, W'^{(c)}_i, b'^{(c)}$).

For the family of optimisation algorithms based on gradient descent, we need to know the gradient of the likelihood with respect to the model parameters, which is given in (8), where $x \in \mathcal{X}$ is a latent entity, and $c \in V$ is an observed predicate (corresponding to the top and bottom rows of figure 3). Note that we extend the definition of energy from situations to entities in the obvious way: half the energy of an entity's links, plus its bias energy. A full derivation of (8) is given in the appendix.

$$\begin{aligned}
\frac{\partial}{\partial\theta} \log P(c) &= \mathbb{E}_{x|c}\left[\frac{\partial}{\partial\theta}\left(-E^b(x)\right)\right] \\
&- \mathbb{E}_{x}\left[\frac{\partial}{\partial\theta}\left(-E^b(x)\right)\right] \\
&+ \mathbb{E}_{x|c}\left[(1 - t_c(x))\,\frac{\partial}{\partial\theta}\left(-E^p(x, c)\right)\right] \\
&- \mathbb{E}_{x|c}\left[\mathbb{E}_{c'|x}\left[(1 - t_{c'}(x))\,\frac{\partial}{\partial\theta}\left(-E^p(x, c')\right)\right]\right]
\end{aligned} \tag{8}$$

There are four terms in this gradient: the first two are for the background distribution, and the last two are for the semantic functions. In both cases, one term is positive, and conditioned on the data, while the other term is negative, and represents the predictions of the model.

Calculating the expectations exactly is infeasible, as this requires summing over all possible configurations. Instead, we can use a Markov Chain Monte Carlo method, as typically done for Latent Dirichlet Allocation (Blei et al., 2003; Griffiths and Steyvers, 2004). Our aim is to sample values of x and c, and use these samples to approximate the expectations: rather than summing over all values, we just consider the samples. For each token in the training data, we introduce a latent entity vector, which we use to approximate the first, third, and fourth terms in (8). Additionally, we introduce a latent predicate for each latent entity, which we use to approximate the fourth term – this latent predicate is analogous to the negative samples used by Mikolov et al. (2013).

When resampling a latent entity conditioned on the data, the conditional distribution $P(x|c)$ is unknown, and calculating it directly requires summing over the whole semantic space. For this reason, we cannot apply Gibbs sampling (as used in LDA), which relies on knowing the conditional distribution. However, if we compare two entities x and x', the normalisation constant cancels out in the ratio $P(x'|c)/P(x|c)$, so we can use the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). Given the current sample x, we can uniformly choose one unit to switch on, and one to switch off, to get a proposal x'. If the ratio of probabilities shown in (9) is above 1, we switch the sample to x'; if it's below 1, it is the probability of switching to x'.

$$\frac{P(x'|c)}{P(x|c)} = \frac{\exp\left(-E^b(x')\right)\frac{1}{Z_{x'}}t_c(x')}{\exp\left(-E^b(x)\right)\frac{1}{Z_x}t_c(x)} \quad (9)$$

Although Metropolis-Hastings avoids the need to calculate the normalisation constant Z of the background distribution, we still have the normalisation constant Z_x of choosing a predicate given an entity. This constant represents the number of predicates true of the entity (weighted by frequency). The intuitive explanation is that we should sample an entity which few predicates are true of, rather than an entity which many predicates are true of. We approximate this constant by assuming that we have an independent contribution from each dimension of x. We first average over all predicates (weighted by frequency), to get the average predicate W^{avg}. We then take the exponential of W^{avg} for the dimensions that we are proposing switching off and on – intuitively, if many predicates have a large weight for a given dimension, then many predicates will be true of an entity where that dimension is active. This is shown in (10), where x and x' differ in dimensions i and i' only, and where k is a constant.

$$\frac{Z_x}{Z_{x'}} \approx \exp\left(k\left(W_i^{avg} - W_{i'}^{avg}\right)\right) \quad (10)$$

We must also resample latent predicates given a latent entity, for the fourth term in (8). This can similarly be done using the Metropolis-Hastings algorithm, according to the ratio shown in (11).

$$\frac{P(c'|x)}{P(c|x)} = \frac{f_{c'}t_{c'}(x)}{f_c t_c(x)} \quad (11)$$

Finally, we need to resample entities from the background distribution, for the second term in (8). Rather than recalculating the samples from scratch after each weight update, we used fantasy particles (persistent Markov chains), which Tieleman (2008) found effective for training RBMs. Resampling a particle can be done more straightforwardly than resampling the latent entities – we can sample each entity conditioned on the other entities in the situation, which can be done exactly using belief propagation (see Yedidia et al. (2003) and references therein), as Swersky et al. (2012) applied to the Cardinality RBM.

To make weight updates from the gradients, we used AdaGrad (Duchi et al., 2011), with exponential decay of the sum of squared gradients. We also used L1 and L2 regularisation, which determines our prior over model parameters.

We found that using a random initialisation is possible, but seems to lead to a long training time, due to slow convergence. We suspect that this could be because the co-occurrence of predicates is mediated via at least two latent vectors, which leads to mixing of semantic classes in each dimension, particularly in the early stages of training. Such behaviour can happen with complicated topic models – for example, Ó Séaghdha (2010) found this for their "Dual Topic" model. One method to reduce convergence time is to initialise predicate parameters using pre-trained vectors. The link parameters can then be initialised

as follows: we consider a situation with just one entity, and for each predicate, we find the mean-field entity vector given the pre-trained predicate parameters; we then fix all entity vectors in our training corpus to be these mean-field vectors, and find the positive pointwise mutual information of each each pair of entity dimensions, for each link label. In particular, we initialised predicate parameters using our sparse SVO Word2Vec vectors, which we describe in section 4.2.

4 Training and Initial Experiments

In this section, we report the first experiments carried out within our framework.

4.1 Training Data

Training our model requires a corpus of DMRS graphs. In particular, we used WikiWoods, an automatically parsed version of the July 2008 dump of the full English Wikipedia, distributed by DELPH-IN[5]. This resource was produced by Flickinger et al. (2010), using the English Resource Grammar (ERG; Flickinger, 2000), trained on the manually treebanked subcorpus WeScience (Ytrestøl et al., 2009), and implemented with the PET parser (Callmeier, 2001; Toutanova et al., 2005). To preprocess the corpus, we used the python packages pydelphin[6] (developed by Michael Goodman), and pydmrs[7] (Copestake et al., 2016).

For simplicity, we restricted attention to subject-verb-object (SVO) triples, although we should stress that this is not an inherent limitation of our model, which could be applied to arbitrary graphs. We searched for all verbs in the WikiWoods treebank, excluding modals, that had either an ARG1 or an ARG2, or both. We kept all instances whose arguments were nominal, excluding pronouns and proper nouns. The ERG does not automatically convert out-of-vocabulary items from their surface form to lemmatised predicates, so we applied WordNet's morphological processor Morphy (Fellbaum, 1998), as available in NLTK (Bird et al., 2009). Finally, we filtered out situations including rare predicates, so that every predicate appears at least five times in the dataset.

As a result of this process, all data was of the form (*verb*, ARG1, ARG2), where one (but not both) of the arguments may be missing. A summary is given in table 1. In total, the dataset contains 72m tokens, with 88,526 distinct predicates.

Situation type	No. instances
Both arguments	10,091,234
ARG1 only	6,301,280
ARG2 only	14,868,213
Total	31,260,727

Table 1: Size of the training data.

4.2 Evaluation

As our first attempt at evaluation, we chose to look at two lexical similarity datasets. The aim of this evaluation was simply to verify that the model was learning something reasonable. We did not expect this task to illustrate our model's strengths, since we need richer tasks to exploit its full expressiveness. Both of our chosen datasets aim to evaluate similarity, rather than thematic relatedness: the first is Hill et al. (2015)'s SimLex-999 dataset, and the second is Finkelstein et al. (2001)'s WordSim-353 dataset, which was split by Agirre et al. (2009) into similarity and relatedness subsets. So far, we have not tuned hyperparameters.

Results are given in table 2. We also trained Mikolov et al. (2013)'s Word2Vec model on the SVO data described in section 4.1, in order to give a direct comparison of models on the same training data. In particular, we used the continuous bag-of-words model with negative sampling, as implemented in Řehůřek and Sojka (2010)'s *gensim* package, with off-the-shelf hyperparameter settings. We also converted these to sparse vectors using Faruqui et al. (2015)'s algorithm, again using off-the-shelf hyperparameter settings. To measure similarity of our semantic functions, we treated each function's parameters as a vector and used cosine similarity, for simplicity.

For comparison, we also include the performance of Word2Vec when trained on raw text. For SimLex-999, we give the results reported by Hill et al. (2015), where the 2-word window model was the best performing model that they tested. For WordSim-353, we trained a model on the full WikiWoods text, after stripping all punctuation and converting to lowercase. We used the *gensim* implementation with off-the-shelf settings, except for window size (2 or 10) and dimension (200, as recommended by Hill et al.). In fact, our re-trained model performed better on SimLex-999 than Hill

[5] http://moin.delph-in.net/WikiWoods
[6] https://github.com/delph-in/pydelphin
[7] https://github.com/delph-in/pydmrs

Model	SimLex Nouns	SimLex Verbs	WordSim Sim.	WordSim Rel.
Word2Vec (10-word window)	.28	.11	**.69**	.46
Word2Vec (2-word window)	.30	.16	.65	.34
SVO Word2Vec	.44	.18	.61	.24
Sparse SVO Word2Vec	**.45**	**.27**	.63	.30
Semantic Functions	.26	.14	.34	**.01**

Table 2: Spearman rank correlation of different models with average annotator judgements. Note that we would like to have a *low* score on the final column (which measures relatedness, rather than similarity).

flood / *water* (related verb and noun)	.06
flood / *water* (related nouns)	.43
law / *lawyer* (related nouns)	.44
sadness / *joy* (near-antonyms)	.77
happiness / *joy* (near-synonyms)	.78
aunt / *uncle* (differ in a single feature)	.90
cat / *dog* (differ in many features)	.92

Table 3: Similarity scores for thematically related words, and various types of co-hyponym.

et al. reported (even when we used less preprocessing or a different edition of Wikipedia), although still worse than our sparse SVO Word2Vec model.

It is interesting to note that training Word2Vec on verbs and their arguments gives noticeably better results on SimLex-999 than training on full sentences, even though far less data is being used: $\sim$72m tokens, rather than $\sim$1000m. The better performance suggests that semantic dependencies may provide more informative contexts than simple word windows. This is in line with previous results, such as Levy and Goldberg (2014a)'s work on using syntactic dependencies. Nonetheless, this result deserves further investigation.

Of all the models we tested, only our semantic function model failed on the relatedness subset of WordSim-353. We take this as a positive result, since it means the model clearly distinguishes relatedness and similarity.

Examples of thematically related predicates and various kinds of co-hyponym are given in table 3, along with our model's similarity scores. However, it is not clear that it is possible, or even desirable, to represent these varied relationships on a single scale of similarity. For example, it could be sensible to treat *aunt* and *uncle* either as synonyms (they refer to relatives of the same degree of relatedness) or as antonyms (they are "opposite" in some sense). Which view is more appropriate will depend on the application, or on the context.

Nouns and verbs are very strongly distinguished, which we would expect given the structure of our model. This can be seen in the similarity scores between *flood* and *water*, when *flood* is considered either as a verb or as a noun.[8] SimLex-999 generally assigns low scores to near-antonyms, and to pairs differing in a single feature, which might explain why the performance of our model is not higher on this task. However, the separation of thematically related predicates from co-hyponyms is a promising result.

5 Related Work

As mentioned above, Coecke et al. (2010) and Baroni et al. (2014) introduce a tensor-based framework that incorporates argument structure through tensor contraction. However, for logical inference, we need to know how one vector can entail another. Grefenstette (2013) explores one method to do this; however, they do not show that this approach is learnable from distributional information, and furthermore, they prove that quantifiers cannot be expressed with tensors.

Balkır (2014), working in the tensorial framework, uses the quantum mechanical notion of a "mixed state" to model uncertainty. However, this doubles the number of tensor indices, so squares the number of dimensions (e.g. vectors become matrices). In the original framework, expressions with several arguments already have a high dimensionality (e.g. *whose* is represented by a fifth-order tensor), and this problem becomes worse.

Vilnis and McCallum (2015) embed predicates as Gaussian distributions over vectors. By assuming covariances are diagonal, this only doubles the number of dimensions (N dimensions for the mean, and N for the covariances). However, similarly to Mikolov et al. (2013), they simply assume

[8] We considered the ERG predicates `_flood_v_cause` and `_flood_n_of`, which were the most frequent predicates in WikiWoods for *flood*, for each part of speech.

that nearby words have similar meanings, so the model does not naturally capture compositionality or argument structure.

In both Balkır's and Vilnis and McCallum's models, they use the probability of a vector given a word – in the notation from section 2.1, $P(x|t)$. However, the opposite conditional probability, $P(t|x)$, more easily allows composition. For instance, if we know two predicates are true (t_1 and t_2), we cannot easily combine $P(x|t_1)$ and $P(x|t_2)$ to get $P(x|t_1, t_2)$ – intuitively, we're generating x twice. In contrast, for semantic functions, we can write $P(t_1, t_2|x) = P(t_1|x)P(t_2|x)$.

Gärdenfors (2004) argues concepts should be modelled as convex subsets of a semantic space. Erk (2009) builds on this idea, but their model requires pre-trained count vectors, while we learn our representations directly. McMahan and Stone (2015) also learn representations directly, considering colour terms, which are grounded in a well-understood perceptual space. Instead of considering a single subset, they use a probability distribution over subsets: $P(A|t)$ for $A \subset \mathcal{X}$. This is more general than a semantic function $P(t|x)$, since we can write $P(t|x) = \sum_{A \ni v} P(A|t)$. However, this framework may be *too* general, since it means we cannot determine the truth of a predicate until we know the entire set A. To avoid this issue, they factorise the distribution, by assuming different boundaries of the set are independent. However, this is equivalent to considering $P(t|x)$ directly, along with some constraints on this function. Indeed, for the experiments they describe, it is sufficient to know a semantic function $P(t|x)$. Furthermore, McMahan and Stone find expressions like *greenish* which are nonconvex in perceptual space, which suggests that representing concepts with convex sets may not be the right way to go.

Our semantic functions are similar to Cooper et al. (2015)'s probabilistic type judgements, which they introduce within the framework of Type Theory with Records (Cooper, 2005), a rich semantic theory. However, one difference between our models is that they represent situations in terms of situation types, while we are careful to define our semantic space without reference to any predicates. More practically, although they outline how their model might be learned, they assume we have access to type judgements for observed situations. In contrast, we describe how a model can be learned from observed utterances, which was necessary for us to train a model on a corpus.

Goodman and Lassiter (2014) propose another linguistically motivated probabilistic model, using the stochastic λ-calculus (more concretely, probabilistic programs written in Church). However, they rely on relatively complex generative processes, specific to individual semantic domains, where each word's meaning may be represented by a complex expression. For a wide-scale system, such structures would need to be extended to cover all concepts. In contrast, our model assumes a direct mapping between predicates and semantic functions, with a relatively simple generative structure determined by semantic dependencies.

Finally, our approach should be distinguished from work which takes pre-trained distributional vectors, and uses them within a richer semantic model. For example, Herbelot and Vecchi (2015) construct a mapping from a distributional vector to judgements of which quantifier is most appropriate for a range of properties. Erk (2016) uses distributional similarity to probabilistically infer properties of one concept, given properties of another. Beltagy et al. (2016) use distributional similarity to produce weighted inference rules, which they incorporate in a Markov Logic Network. Unlike these authors, we aim to directly learn interpretable representations, rather than interpret given representations.

6 Conclusion

We have introduced a novel framework for distributional semantics, where each predicate is represented as a function, expressing how applicable the predicate is to different entities. We have shown how this approach can capture semantic phenomena which are challenging for standard vector space models. We have explained how our framework can be implemented, and trained on a corpus of DMRS graphs. Finally, our initial evaluation on similarity datasets demonstrates the feasibility of this approach, and shows that thematically related words are not given similar representations. In future work, we plan to use richer tasks which exploit the model's expressiveness.

Acknowledgments

This work was funded by a Schiff Foundation Studentship. We would also like to thank Yarin Gal, who gave useful feedback on the specification of our generative model.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of the 2009 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27.

Keith Allan. 2001. *Natural Language Semantics*. Blackwell Publishers.

Esma Balkır. 2014. Using density matrices in a compositional distributional model of meaning. Master's thesis, University of Oxford.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2012. Abstract meaning representation (amr) 1.0 specification. In *Proceedings of the 11th Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology*, 9.

Jon Barwise and John Perry. 1983. *Situations and Attitudes*. MIT Press.

Islam Beltagy, Stephen Roller, Pengxiang Cheng, Katrin Erk, and Raymond J. Mooney. 2016. Representing meaning with a combination of logical and distributional models.

Maxwell R Bennett and Peter Michael Stephan Hacker. 2008. *History of cognitive neuroscience*. John Wiley & Sons.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly Media, Inc.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *the Journal of Machine Learning Research*, 3:993–1022.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The Prague dependency treebank. In *Treebanks*, pages 103–127. Springer.

Ulrich Callmeier. 2001. Efficient parsing with large-scale unification grammars. Master's thesis, Universität des Saarlandes, Saarbrücken, Germany.

Ronnie Cann. 1993. *Formal semantics: An introduction*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36:345–384.

Robin Cooper, Simon Dobnik, Staffan Larsson, and Shalom Lappin. 2015. Probabilistic type theory and natural language semantics. *LiLT (Linguistic Issues in Language Technology)*, 10.

Robin Cooper. 2005. Austinian truth, attitudes and type theory. *Research on Language and Computation*, 3(2-3):333–362.

Ann Copestake and Aurelie Herbelot. 2012. Lexicalised compositionality. Unpublished draft.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. Minimal Recursion Semantics: An introduction. *Research on Language and Computation*.

Ann Copestake, Guy Emerson, Michael Wayne Goodman, Matic Horvat, Alexander Kuhnle, and Ewa Muszyńska. 2016. Resources for building applications with Dependency Minimal Recursion Semantics. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Ann Copestake. 2009. Slacker semantics: Why superficiality, dependency and avoidance of commitment can be the right way to go. In *Proceedings of 12th Conference of the European Chapter of the Association for Computational Linguistics*.

Donald Davidson. 1967. The logical form of action sentences. In Nicholas Rescher, editor, *The Logic of Decision and Action*, chapter 3, pages 81–95. University of Pittsburgh Press.

Georgiana Dinu, Stefan Thater, and Sören Laue. 2012. A comparison of models of word meaning in context. In *Proceedings of the 13th Conference of the North American Chapter of the Association for Computational Linguistics*, pages 611–615.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

Michael Dummett. 1976. What is a theory of meaning? (II). In Gareth Evans and John McDowell, editors, *Truth and Meaning*, pages 67–137. Clarendon Press (Oxford).

Michael Dummett. 1978. *What Do I Know When I Know a Language?* Stockholm University. Reprinted in Dummett (1993) *Seas of Language*, pages 94–105.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 13th Conference on Empirical Methods in Natural Language Processing*, pages 897–906. Association for Computational Linguistics.

Katrin Erk. 2009. Representing words as regions in vector space. In *Proceedings of the 13th Conference on Computational Natural Language Learning*, pages 57–65. Association for Computational Linguistics.

Katrin Erk. 2016. What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics*, 9(17):1–63.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah Smith. 2015. Sparse overcomplete word vector representations. In *Proceedings of the 53rd Annual Conference of the Association for Computational Linguistics*.

Christiane Fellbaum. 1998. *WordNet*. Blackwell Publishers.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on the World Wide Web*, pages 406–414. Association for Computing Machinery.

Dan Flickinger, Stephan Oepen, and Gisle Ytrestøl. 2010. WikiWoods: Syntacto-semantic annotation for English Wikipedia. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*.

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*.

Peter Gärdenfors. 2004. *Conceptual spaces: The geometry of thought*. MIT Press, second edition.

Dan Garrette, Katrin Erk, and Raymond Mooney. 2011. Integrating logical representations with probabilistic information using Markov logic. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, pages 105–114. Association for Computational Linguistics.

Noah D Goodman and Daniel Lassiter. 2014. Probabilistic semantics and pragmatics: Uncertainty in language and thought. *Handbook of Contemporary Semantic Theory*.

Edward Grefenstette. 2013. Towards a formal distributional semantics: Simulating logical calculi with tensors. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

W. Keith Hastings. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: Mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 22–32. Association for Computational Linguistics.

Aurélie Herbelot. 2015. Mr Darcy and Mr Toad, gentlemen: distributional names and their kinds. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 151–161.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

Hans Kamp and Uwe Reyle. 1993. From discourse to logic; introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory.

Anthony Kenny. 2010. Concepts, brains, and behaviour. *Grazer Philosophische Studien*, 81(1):105–113.

William Labov. 1973. The boundaries of words and their meanings. In Charles-James N. Bailey and Roger W. Shuy, editors, *New ways of analyzing variation in English*, pages 340–73. Georgetown University Press.

Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*.

Omer Levy and Yoav Goldberg. 2014a. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 302–308.

Omer Levy and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc.

Godehard Link. 2002. The logical analysis of plurals and mass terms: A lattice-theoretical approach. In Paul Portner and Barbara H. Partee, editors, *Formal semantics: The essential readings*, chapter 4, pages 127–146. Blackwell Publishers.

Michael E McCloskey and Sam Glucksberg. 1978. Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, 6(4):462–472.

Brian McMahan and Matthew Stone. 2015. A Bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*, 3:103–115.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations*.

Brian Murphy, Partha Pratim Talukdar, and Tom M Mitchell. 2012. Learning effective and interpretable semantic models using non-negative sparse embedding. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1933–1950. Association for Computational Linguistics.

Gregory Leo Murphy. 2002. *The Big Book of Concepts*. MIT Press.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444. Association for Computational Linguistics.

Terence Parsons. 1990. *Events in the Semantics of English: A Study in Subatomic Semantics*. Current Studies in Linguistics. MIT Press.

Willard Van Orman Quine. 1960. *Word and Object*. MIT Press.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA.

Matthew Richardson and Pedro Domingos. 2006. Markov logic networks. *Machine learning*, 62(1-2):107–136.

Paul Smolensky. 1986. Information processing in dynamical systems: Foundations of harmony theory. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, pages 194–281. MIT Press.

Kevin Swersky, Ilya Sutskever, Daniel Tarlow, Richard S Zemel, Ruslan R Salakhutdinov, and Ryan P Adams. 2012. Cardinality Restricted Boltzmann Machines. In *Advances in Neural Information Processing Systems*, pages 3293–3301.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2011. Word meaning in context: A simple and effective vector model. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 1134–1143.

Tijmen Tieleman. 2008. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1064–1071. Association for Computing Machinery.

Kristina Toutanova, Christoper D. Manning, Dan Flickinger, and Stephan Oepen. 2005. Stochastic HPSG parse selection using the Redwoods corpus. *Journal of Research on Language and Computation*.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*.

Luke Vilnis and Andrew McCallum. 2015. Word representations via Gaussian embedding. In *Proceedings of the 3rd International Conference on Learning Representations*.

Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. 2003. Understanding Belief Propagation and its generalizations. In *Exploring Artificial Intelligence in the New Millennium*, chapter 8, pages 239–269.

Gisle Ytrestøl, Stephan Oepen, and Daniel Flickinger. 2009. Extracting and annotating Wikipedia subdomains. In *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories*.

Appendix: Derivation of Gradients

In this section, we derive equation (8). As our model generates predicates from entities, to find the probability of observing the predicates, we need to sum over all possible entities. After then applying the chain rule to log, and expanding $P(x, c)$, we obtain the expression below.

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\log P(c) = \frac{\partial}{\partial\theta}\log\sum_x P(x,c) \\
&= \frac{\frac{\partial}{\partial\theta}\sum_x P(x,c)}{\sum_{x'} P(x',c)} \\
&= \frac{\frac{\partial}{\partial\theta}\sum_x \frac{1}{Z_x} f_c t_c(x) \frac{1}{Z}\exp\left(-E^b(x)\right)}{\sum_{x'} P(x',c)}
\end{aligned}$$

When we now apply the product rule, we will get four terms, but we can make use of the fact that the derivatives of all four terms are multiples of the original term:

$$\frac{\partial}{\partial\theta}e^{-E^b(x)} = e^{-E^b(x)}\frac{\partial}{\partial\theta}\left(-E^b(x)\right)$$

$$\frac{\partial}{\partial\theta}t_c(x) = t_c(x)\left(1 - t_c(x)\right)\frac{\partial}{\partial\theta}\left(-E^p(x,c)\right)$$

$$\frac{\partial}{\partial\theta}\frac{1}{Z_x} = \frac{-1}{Z_x^2}\frac{\partial}{\partial\theta}Z_x$$

$$\frac{\partial}{\partial\theta}\frac{1}{Z} = \frac{-1}{Z^2}\frac{\partial}{\partial\theta}Z$$

This allows us to derive:

$$
\begin{aligned}
=&\sum_x \frac{P(x,c)}{\sum_{x'} P(x',c)} \left[\frac{\partial}{\partial\theta}\left(-E^b(x)\right) \right. \\
&\qquad + (1-t_c(x)) \frac{\partial}{\partial\theta}\left(-E^p(x,c)\right) \\
&\qquad \left. - \frac{1}{Z_x}\frac{\partial}{\partial\theta} Z_x \right] \\
&- \frac{\sum_x P(x,c)}{\sum_{x'} P(x',c)} \frac{1}{Z}\frac{\partial}{\partial\theta} Z
\end{aligned}
$$

We can now simplify using conditional probabilities, and expand the derivatives of the normalisation constants:

$$
\begin{aligned}
=&\sum_x P(x|c) \left[\frac{\partial}{\partial\theta}\left(-E^b(x)\right) \right. \\
&\qquad + (1-t_c(x)) \frac{\partial}{\partial\theta}\left(-E^p(x,c)\right) \\
&\qquad \left. - \frac{1}{Z_x}\frac{\partial}{\partial\theta} \sum_{c'} f_{c'} t_{c'}(x) \right] \\
&- \frac{1}{Z}\frac{\partial}{\partial\theta} \sum_x \exp\left(-E^b(x)\right)
\end{aligned}
$$

$$
\begin{aligned}
=&\sum_x P(x|c) \left[\frac{\partial}{\partial\theta}\left(-E^b(x)\right) \right. \\
&\quad + (1-t_c(x)) \frac{\partial}{\partial\theta}\left(-E^p(x,c)\right) \\
&\quad \left. - \sum_{c'} \frac{f_{c'} t_{c'}(x)}{Z_x} (1-t_{c'}(x)) \frac{\partial}{\partial\theta}\left(-E^p(x,c')\right) \right] \\
&- \sum_x \frac{\exp\left(-E^b(x)\right)}{Z} \frac{\partial}{\partial\theta}\left(-E^b(x)\right)
\end{aligned}
$$

$$
\begin{aligned}
=&\sum_x P(x|c) \left[\frac{\partial}{\partial\theta}\left(-E^b(x)\right) \right. \\
&\quad + (1-t_c(x)) \frac{\partial}{\partial\theta}\left(-E^p(x,c)\right) \\
&\quad \left. - \sum_{c'} P(c'|x) (1-t_{c'}(x)) \frac{\partial}{\partial\theta}\left(-E^p(x,c')\right) \right] \\
&- \sum_x P(x) \frac{\partial}{\partial\theta}\left(-E^b(x)\right)
\end{aligned}
$$

Finally, we write expectations instead of sums of probabilities:

$$
\begin{aligned}
=&\mathbb{E}_{x|c} \left[\frac{\partial}{\partial\theta}\left(-E^b(x)\right) \right. \\
&\qquad + (1-t_c(x)) \frac{\partial}{\partial\theta}\left(-E^p(x,c)\right) \\
&\qquad \left. - \mathbb{E}_{c'|x} \left[(1-t_{c'}(x)) \frac{\partial}{\partial\theta}\left(-E^p(x,c')\right) \right] \right] \\
&- \mathbb{E}_x \left[\frac{\partial}{\partial\theta}\left(-E^b(x)\right) \right]
\end{aligned}
$$