# Compositional semantics in a probabilistic framework

Guy Emerson

## Introduction and Overview

The most popular approaches to distributional semantics represent meanings as points in a vector space, either as *count* vectors (e.g. Turney and Pantel (2010)) or as embedding vectors (e.g. Mikolov et al. (2013)). However, vectors do not provide 'natural' composition operations that have clear analogues with operations in formal semantics. 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 quantifiers.

Here, we present an probabilistic framework for distributional semantics, which gives a natural account of various semantic phenomena, such as vagueness, context-dependence, and quantification. We further show that quantification in classical logic reduces to a special case.

## Semantic Functions

We can separate entities from the predicates describing them; intuitively, the world is the same however we describe it. For any entity, we can ask how applicable each predicate is to the entity. More formally, if we take entities to be points in some semantic space $S$ (whose dimensions may denote different features), then we can take the meaning of a predicate to be a function from $S$ to values in the interval $[0, 1]$, denoting how likely a competent speaker is to judge the predicate applicable to the entity (or 'true' of the entity).

Representing predicates as functions allows us to naturally capture vagueness (a predicate can be equally applicable to a region of space), and using values between 0 and 1 allows us to naturally capture gradedness (a predicate can be more applicable to some points). Representing a predicate as a region of space is not a new idea, and our proposal is in the spirit of Gärdenfors (2004), Erk (2009), Vilnis and McCallum (2015), Balkır (2014), and McMahan and Stone (2015). It is also in the spirit of the probabilisitic models proposed by Cooper et al. (2015) and Goodman and Lassiter (2014). Compared to each of these proposals, we believe our model has either a better defined compositional semantics, or a better way of learning from corpus data.

## Incorporating Semantic Functions in Minimal Recursion Semantics

Using Dependency Minimal Recursion Semantics (DMRS; Copestake (2009)), we can represent the meaning of a linguistic expression as a directed graph: nodes represent predicates/entities (relying on a 1:1 correspondence between them) and edges represent argument structure and scopal constraints.

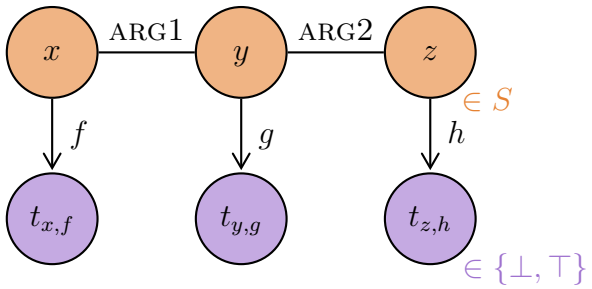


Figure 1: A graph with three predicates

As mentioned above, we can separate the nodes into two parts, representing predicates and entities. This allows us to define a probabilistic graphical model, with an example given in figure 1. This might represent a transitive verb (in the middle), and its two arguments (subject on the left and object on the right). The node $y$ represents the event that the verb describes, while the node $t_{y,g}$ represents whether or not the verb accurately describes the event. We can view these nodes as together representing a situation, in the sense of Barwise and Perry (1983). The remaining nodes similarly represent the two individuals involved in the event, and whether or not the nouns accurately describe them. 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 $P(x, y)\, P(y, z)$. The conditional distributions for the predicate nodes are defined in terms of the semantic functions, e.g. $P(t_{x,f} = \top | x) = f(x)$.

The probability of the entities taking the specific values $x$, $y$, $z$, and all of the predicates being true, is then given by $P(x, y, z)f(x)g(y)h(z)$. The first term, in orange, can be thought of as situational knowledge or world knowledge – what entities are likely to occur together? The second term, in purple, can be thought of as lexical semantic knowledge – what entities are different predicates applicable to?

This separation of entities from predicates allows us to naturally capture context-dependent meanings. Following the terminology of Quine (1960), we can separate an expression's context-independent *standing* meaning from its 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 all the predicates in the graph – this represents its occasion meaning. Unlike previous approaches, such as Dinu et al. (2012), Erk and Padó (2008), and Thater et al. (2011), meanings in and out of context are represented by different kinds of objects, reflecting a type/token distinction.

## Machine Learning

In order to train such a model, it is necessary to explicitly parametrise the background distribution and semantic functions, so that we can optimise these parameters, given a corpus. For simplicity, we can take the semantic space $S$ to be binary-valued vectors, $\{0, 1\}^N$. We can then define the background distributon using a Restricted Boltzmann Machine – intuitively, for any link in a DMRS graph, the Boltzmann Machine can represent soft constraints on which kinds of entities are related by such a link. We can represent the semantic functions by feedforward neural networks – intuitively, these represent which features of the entities are relevant to a predicate. The background distribution and the semantic functions can together be used to define a generative process that takes a DMRS graph without predicates and generates a predicate for each node. Given a corpus of DMRS graphs (such as WikiWoods Flickinger et al. (2010)), we can train the model by maximising the probability of generating the training data.

## Interpretation of Quantifiers

Given a prior distribution over situations (world knowledge), and given semantic functions (lexical semantic knowledge), we aim to define a probabilistic truth value for a sentence. Unlike Herbelot and Vecchi (2015), we are not mapping from a distributional space to a model structure, but directly interpreting quantifiers in our model.

As a graphical model, there are directed links from entities to predicates – i.e. the distribution over situations is not affected by the truth we assign to predicates. Similarly, we can now try to define the truth of quantified expressions so that the underlying distribution is not affected. However, unlike the predicates, the distributions for quantifiers will be functions on the joint distribution, rather than functions on the values of the random variables.

Given a scoping (so one quantifier per entity), we define one random variable per quantifier, each taking a binary value. We define their distributions recursively, bottom up through the scoping. Each quantifier depends on its restriction and body, which may either be a predicate, or a quantifier – both are binary-valued random variables, and both are associated with one entity. How the two random variables are combined depends on the particular quantifier.

It is possible to define composition functions for universal and existential quantifiers in such a way that classical predicate logic with a model theory can be derived as a special case, where semantic functions only take the values 0 or 1 (so that they define a denotation), and where the background distribution corresponds to picking an entity from the model. However, our framework can also go further, naturally capturing uncertain world knowledge, uncertain lexical semantic knowledge, and "fuzzy" quantifiers, such as in generic statements.

## References


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

Baroni, M., R. Bernardi, and R. Zamparelli (2014). Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology 9*.

Barwise, J. and J. Perry (1983). *Situations and Attitudes*. MIT Press.

Coecke, B., M. Sadrzadeh, and S. Clark (2010). Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis 36*, 345–384.

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

Copestake, A. (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*.

Dinu, G., S. Thater, and S. 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*, pp. 611–615.

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

Erk, K. and S. 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*, pp. 897–906. Association for Computational Linguistics.

Flickinger, D., S. Oepen, and G. Ytrestøl (2010). WikiWoods: Syntacto-semantic annotation for English Wikipedia. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought* (Second ed.). MIT press.

Goodman, N. D. and D. Lassiter (2014). Probabilistic semantics and pragmatics: Uncertainty in language and thought. *Handbook of Contemporary Semantic Theory. Wiley-Blackwell*.

Herbelot, A. and E. M. Vecchi (2015). Building a shared world: Mapping distributional to model-theoretic semantic spaces. In *Proceeedings of the 20th Conference on Empirical Methods in Natural Language Processing*.

McMahan, B. and M. Stone (2015). A Bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics 3*, 103–115.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations*.

Quine, W. V. O. (1960). *Word and Object*. MIT Press.

Thater, S., H. Fürstenau, and M. Pinkal (2011). Word meaning in context: A simple and effective vector model. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pp. 1134–1143.

Turney, P. D. and P. Pantel (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*.

Vilnis, L. and A. McCallum (2015). Word representations via Gaussian embedding. In *Proceedings of the 3rd International Conference on Learning Representations*.