What are the Goals of Distributional Semantics?

Guy Emerson
Department of Computer Science and Technology
University of Cambridge
gete2@cam.ac.uk

Abstract

Distributional semantic models have become a mainstay in NLP, providing useful features for downstream tasks. However, assessing long-term progress requires explicit long-term goals. In this paper, I take a broad linguistic perspective, looking at how well current models can deal with various semantic challenges. Given stark differences between models proposed in different subfields, a broad perspective is needed to see how we could integrate them. I conclude that, while linguistic insights can guide the design of model architectures, future progress will require balancing the often conflicting demands of linguistic expressiveness and computational tractability.

1 Introduction

In order to assess progress in any field, the goals need to be clear. In assessing progress in semantics, Koller (2016) contrasts "top-down" and "bottom-up" approaches: a top-down approach begins with an overarching goal, and tries to build a model to reach it; a bottom-up approach begins with existing models, and tries to extend them towards new goals.[1] Like much of NLP, distributional semantics is largely bottom-up: the goals are usually to improve performance on particular tasks, or particular datasets. Aiming to improve NLP applications is of course a legitimate decision, but Koller points out a problem if there is no top-down goal: "Bottom-up theories are intrinsically unfalsifiable... We won't know where distributional semantics is going until it has a top-down element". This is contrasted against truth-conditional semantics, a traditional linguistic approach which is largely top-down: "truth-conditional semantics hasn't reached its goal, but at least we knew what the goal was".

In this paper, I take a long-term linguistic perspective, where the top-down goal is to characterise the meanings of all utterances in a language. This is an ambitious goal, and a broad one. To make this goal more precise, in the following sections I will elaborate on several aspects of meaning which could be considered crucial. For each aspect, I identify a plausible goal, lay out out the space of possible models, place existing work in this space, and evaluate which approaches seem most promising. By making the goals explicit, we can assess whether we are heading in the right direction, and we can assess what still needs to be done. If a reader should disagree with my conclusions, they should start by looking at my goals.

[1]For further discussion, see: Bender and Koller (2020).

2 Background: Distributional Semantics

The aim of distributional semantics is to learn the meanings of linguistic expressions from a corpus of text. The core idea, known as the *distributional hypothesis*, is that the contexts in which an expression appears give us information about its meaning.[2]

The idea has roots in American structuralism (Harris, 1954) and British lexicology (Firth, 1951, 1957)[3], and with the advent of modern computing, it began to be used in practice. In a notable early work, Spärck-Jones (1964) represented word meanings as boolean vectors, based on a thesaurus.

Distributional semantics has become widespread in NLP, first with the rise of count vectors (for an overview, see: Erk, 2012; Clark, 2015), then of word embeddings (Mikolov et al., 2013), and most recently, of contextualised embeddings (Peters et al., 2018; Devlin et al., 2019).[4] What all of these approaches share is that they learn representations in an unsupervised manner on a corpus.

[2]The hypothesis is often stated more narrowly, to say that similar words appear in similar contexts, but in this paper I am interested in semantics beyond just similarity.

[3]Firth used the term *collocational*, not *distributional*.

[4]For connections between count vectors and embeddings, see: Levy and Goldberg (2014); Cotterell et al. (2017); for connections with contextual embeddings: Kong et al. (2020).

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7436–7453
July 5 - 10, 2020.

While much work takes a bottom-up approach, as Koller observes, a notable exception is the type-driven tensorial framework of Coecke et al. (2010) and Baroni et al. (2014), which has broad linguistic goals, and will be mentioned in several sections below. This framework represents the meanings of words as tensors, and constructs phrase meanings using tensor contraction based on predicate-argument structure. For example, there is one vector space for nouns, and a second vector space for sentences, so intransitive verbs are matrices (mapping noun vectors to sentence vectors).

3 Meaning and the World

Language is always *about* something. In this section, I discuss challenges in connecting a semantic model to things in the world.

3.1 Grounding

As Harnad (1990) discusses, if the meanings of words are defined only in terms of other words, these definitions are circular. One goal for a semantic model is to capture how language relates to the world, including sensory perception and motor control – this process of connecting language to the world is called *grounding*.[5]

A purely distributional model is not grounded, as it is only trained on text, with no direct link to the world. There are several ways we could try to ground a distributional model (for an overview, see: Baroni, 2016). The simplest way is to train a distributional model as normal, then combine it with a grounded model. For example, Bruni et al. (2011) concatenate distributional vectors and image feature vectors. This has also been applied to other senses: Kiela et al. (2015) use olfactory data, and Kiela and Clark (2017) use both visual and auditory data. However, while there is grounded information in the sensory dimensions, concatenation leaves the distributional dimensions ungrounded.

A second approach is to find correlations between distributional and sensory features. For example, Bruni et al. (2014) perform SVD on concatenated vectors, Silberer and Lapata (2014) train an autoencoder on concatenated vectors, and Lazaridou et al. (2014) and Bulat et al. (2016) learn a mapping from distributional vectors to visual vectors (and vice versa). However, there is no guarantee that every distributional feature will correlate with sensory features. Distributional features without correlations will remain ungrounded.

Finally, a third approach is joint learning – we define a single model, whose parameters are learnt based on both corpus data and grounded data. For example, Feng and Lapata (2010) train an LDA model (Blei et al., 2003) for both words and "visual words" (clusters of visual features). Lazaridou et al. (2015) use a Skip-gram model (Mikolov et al., 2013) to jointly predict both words and images. Kiros et al. (2014) embed both text and images in a single space, training an RNN to process captions, and a CNN to process images. Pure distributional models look for word co-occurrence patterns, while joint models prefer co-occurrence patterns that match the grounded data. For this reason, I believe joint learning is the right approach to ground corpus data – semantic representations can be connected to grounded data from the outset, rather than trying to make such connections after the fact.

However, we must still make sure that all distributional features are grounded. With Feng and Lapata's LDA model, some topics might only generate words rather than "visual words". Similarly, with Lazaridou et al.'s joint Skip-gram model, some embeddings might only predict words rather than images. Conversely, we also need to make sure that we make full use of corpus data, rather than discarding what is difficult to ground. For example, Kiros et al.'s joint embedding model learns sentence embeddings in order to match them to images. It is not obvious how this approach could be extended so that we can learn embeddings for sentences that cannot be easily depicted in an image.

This leads to the question: how should a joint architecture be designed, so that we can fully learn from corpus data, while ensuring that representations are fully grounded? Grounding is hard, and indeed Kuhnle et al. (2018) find that some semantic constructions (such as superlatives) are much harder for grounded models to learn than others. In the following section, I discuss how language relates to the world. Clarifying this relationship should help us to design good joint architectures.

3.2 Concepts and Referents

How do meanings relate to the world? In truth-conditional semantics, the answer is that meaning is defined in terms of *truth*.[6] If an agent under-

[5] This includes connecting abstract concepts to the world, although such connections are necessarily more indirect. For further discussion, see: Blondin-Massé et al. (2008); Pecher et al. (2011); Pulvermüller (2013); Barsalou et al. (2018)

[6] For a discussion of this point, see: Lewis (1970). For an

stands a language, then in any given situation, they know how to evaluate whether a sentence is true or false of that situation.[7] An advantage of this approach is that it supports logical reasoning, which I will discuss in §5.2. One goal for a semantic theory is to be able to generalise to new situations. This is difficult for traditional truth-conditional semantics, with classical theories challenged on both philosophical grounds (for example: Wittgenstein, 1953, §66–71) and empirical grounds (for example: Rosch, 1975, 1978). However, a machine learning approach seems promising, since generalising to new data is a central aim of machine learning.

For a semantic model to be compatible with truth-conditional semantics, it is necessary to distinguish a *concept* (the meaning of a word) from a *referent* (an entity the word can refer to).[8] The importance of this distinction has been noted for some time (for example: Ogden and Richards, 1923). A concept's set of referents is called its *extension*.[9]

Even if we can construct grounded concept vectors, as discussed in §3.1, there is still the question of how to relate a concept vector to its referents.[10] One option is to embed both concepts and entities in the same space. We then need a way to decide how close the vectors need to be, for the entity to be in the concept's extension. A second option is to embed concepts and referents in distinct spaces. We then need a way to relate the two spaces.

In both cases, we need additional structure beyond representing concepts and referents as points. One solution is to represent a concept by a *region* of space (Gärdenfors, 2000, 2014). Entities embedded inside the region are referents, while those outside are not. For example, McMahan and Stone (2015) learn representations of colour terms, which are grounded in a well-understood perceptual space.

A related idea is to represent a concept as a binary classifier, where an entity is the input.[11] One class is the concept's extension, and the other class is everything else. Larsson (2013) represents the meaning of a perceptual concept as a classifier of perceptual input. A number of authors have trained image classifiers using captioned images (for example: Schlangen et al., 2016; Zarrieß and Schlangen, 2017a,b; Utescher, 2019; Matsson et al., 2019).

Such representations have however seen limited use in distributional semantics. Erk (2009a,b) and Dong et al. (2018) learn regions, but relying on pre-trained vectors, which may have already lost referential information (such as co-reference) that we would like to capture. Jameel and Schockaert (2017) learn a hybrid model, where each word is represented by a point (as a target word) and a region (as a context word). In my own work, I have learnt classifiers (Emerson and Copestake, 2016, 2017a,b), but with a computationally expensive model that is difficult to train. The computational challenge is partially resolved in my most recent work (Emerson, 2020a), but there is still work to be done in scaling up the model to make full use of the corpus data. The best way to design such a model, so that it can both make full use of the data and can be trained efficiently, is an open question.

4 Lexical Meaning

In this section, I discuss challenges in representing the meanings of individual words.

4.1 Vagueness

Entities often fall along a continuum without a sharp cutoff between concepts. This is called *vagueness* (or *gradedness*). (For an overview, see: Sutton, 2013, chapter 1; Van Deemter, 2010.) For example, Labov (1973) investigated the boundaries between concepts like *cup*, *mug*, and *bowl*, asking participants to name drawings of objects. For typical referents, terms were used consistently; meanwhile, for objects that were intermediate between concepts (for example, something wide for a cup but narrow for a bowl), terms were used inconsistently. For these borderline cases, a single person may make different judgements at different times (McCloskey and Glucksberg, 1978). One goal for a semantic model is to capture how it can be unclear whether an entity is an referent of a concept.

One approach is to use *fuzzy* truth values, which are not binary true/false, but rather values in the range [0,1], where 0 is definitely false, 1 is definitely true, and intermediate values represent borderline cases (Zadeh, 1965, 1975). Fuzzy logic has

introduction to truth-conditional semantics, see: Cann (1993); Allan (2001); Kamp and Reyle (2013).

[7] On the notion of *situation*, see: Barwise and Perry (1983). On knowing *how* to evaluate truth values vs. actually evaluating truth values, see: Dummett (1976, 1978).

[8] Following Murphy (2002, pp. 4–5), I use the term *concept* without committing to a particular theory of concepts.

[9] Or *denotation*. In psychology, the term *category* is also used (for example: Smith and Medin, 1981; Murphy, 2002).

[10] While distributional representations can be learnt for named entities (for example: Herbelot, 2015; Boleda et al., 2017), most real-world entities are not mentioned in text.

[11] For deterministic regions and classifiers, there is a one-to-one mapping between them, but this is not true for probabilistic regions and classifiers, due to covariance.

not seen much use in computational linguistics.[12]

A second solution is to stick with binary truth values, but using probability theory to formalise uncertainty about truth, as has been proposed in formal semantics (for example: Lassiter, 2011; Fernández and Larsson, 2014; Sutton, 2015, 2017). At the level of a single concept, there is not much to decide between fuzzy and probabilistic accounts, since both assign values in the range [0,1]. However, we will see in §5.2 that they behave differently at the level of sentences.

Uncertainty has also been incorporated into distributional vector space models. Vilnis and McCallum (2015) extend Mikolov et al.'s Skip-gram model, representing meanings as Gaussian distributions over vectors. Barkan (2017) incorporate uncertainty into Skip-gram using Bayesian inference – rather than optimising word vectors, the aim is to calculate the posterior distribution over word vectors, given the observed data. The posterior is approximated as a Gaussian, so these two approaches produce the same kind of object. Balkır (2014), working within the type-driven tensorial framework (see §2), uses a quantum mechanical "mixed state" to model uncertainty in a tensor. For example, this replaces vectors by matrices, and replaces matrices by fourth-order tensors.

While these approaches represent uncertainty, it is challenging to use them to capture vagueness. This basic problem is this: a distribution allows us to *generate* referents of a concept, but how can we go in the other direction, to *recognise* referents of a concept? It is tempting to classify a point using the probability density at that point, but if we compare a more general term with a more specific term (like *animal* and *dog*), we find a problem: a more general term has its probability mass spread more thinly, and hence has a lower probability density than the more specific term, even if both terms could be considered true. I argued in §3.2 that, to talk about truth, we need to represent predicates as regions of space or as classifiers. While a distribution over a space might at first sight look like a region of space, normalising the probability mass to sum to 1 makes a distribution a different kind of object.

[12] Carvalho et al. (2012) survey fuzzy logic in NLP, noting that its use is in decline, but they do not mention distributional semantics. Proposals such as Monte Carlo Semantics (Bergmair, 2010) and Fuzzy Natural Logic (Novák, 2017) do not provide an approach to distributional semantics. A rare exception is Runkler (2016), who infers fuzzy membership functions from pre-trained vectors.

4.2 Polysemy

The meaning of a word can often be broken up into distinct *senses*. Related senses are called *polysemous*: for example, *school* can refer to a building or an institution. In contrast, *homonymous* senses are unrelated: for example, a *school* of fish. All of the above senses of *school* are also *lexicalised* – established uses that a speaker would have committed to memory, rather than inferring from context. I will discuss context-dependent meaning in §5.3, and focus here on lexicalised meaning. One goal for a semantic model is to capture how a word can have a range of polysemous senses.

One solution is to learn a separate representation for each sense (for example: Schütze, 1998; Rapp, 2004; Li and Jurafsky, 2015; for a survey, see: Camacho-Collados and Pilehvar, 2018). However, deciding on a discrete set of senses is difficult, and practical efforts at compiling dictionaries have not provided a solution. Indeed, the lexicographer Sue Atkins bluntly stated, "I don't believe in word senses".[13] Although the sense of a word varies across usages, there are many ways that we could cluster usages into a discrete set of senses, a point made by many authors (for example: Spärck-Jones, 1964; Kilgarriff, 1997, 2007; Hanks, 2000; Erk, 2010). To quantify this intuition, Erk et al. (2009, 2013) produced the WSsim and Usim datasets, where annotators judged the similarity between dictionary senses, and the similarity between individual usages, respectively. McCarthy et al. (2016) quantify "clusterability" in USim, showing that for some words, usages cannot be clustered into discrete senses. A good semantic model should therefore be able to capture variation in meaning without resorting to finite sense inventories.

We could instead learn a single representation for all polysemous senses together. Indeed, Ruhl (1989) argues that even frequent terms with many apparent senses, such as *bear* and *hit*, can be analysed as having a single underspecified meaning, with the apparent diversity of senses explainable from context. The challenge is then to represent such a meaning without overgeneralising to cases where the word wouldn't be used, and to model how meanings are specialised in context. The second half of this challenge will be discussed in §5.3.

I have already argued in previous sections that we should move away from representing each word as a single vector. As discussed in §4.1, words

[13] Kilgarriff (1997) and Hanks (2000) both quote Atkins.

can be represented with distributions, and such an approach has also been applied to modelling word senses. For example, Athiwaratkun and Wilson (2017) use a mixture of Gaussians, extending Vilnis and McCallum's model to allow multiple senses. However, this ultimately models a fixed number of senses (one for each Gaussian). In principle, a distribution could be parametrised in a more general way, moving beyond finite mixture models. In the type-driven tensorial framework (see §2), Piedeleu et al. (2015) use mixed quantum states, similarly to Balkır's approach (see §4.1). Although they only propose this approach for homonymy, it could plausibly be used for polysemy as well.

If a word is represented by a region, or by a classifier, we don't have the problem of finite sense inventories, as long as the region or classifier is parametrised in a general enough way – for example, a multi-layer neural net classifier, rather than a finite mixture of simple classifiers.

4.3 Hyponymy

In the previous two sections, I discussed meanings of single words. However, words do not exist on their own, and one goal for semantic model is to represent relations between them. A classic relation is *hyponymy*,[14] which describes when one term (the *hyperonym* or *hypernym*) has a more general meaning than another (the *hyponym*). Words that share a hyperonym are called *co-hyponyms*.

In a vector space model, it is not clear how to say if one vector is more general than another. One idea is that a hyperonym should occur in all the contexts of its hyponyms. This is known as the *Distributional Inclusion Hypothesis* (DIH; Weeds et al., 2004; Geffet and Dagan, 2005). Using this idea and tools from information retrieval, Kotlerman et al. (2009, 2010) define the "balAPinc" measure of hyponymy. Herbelot and Ganesalingam (2013) view a vector as a distribution over contexts, using KL-divergence to measure hyponymy. Rei (2013) gives an overview of hyponymy measures, and proposes a weighted cosine measure. For embeddings, the motivation for such measures is less direct, but dimensions can be seen as combinations of contexts. Indeed, Rei and Briscoe (2014) find embeddings perform almost as well as count vectors.

However, a speaker is likely to choose an expression with a degree of generality appropriate for the discourse (the Maxim of Quantity; Grice, 1967), and hence the DIH can be questioned. Rimell (2014) points out that some contexts are highly specific. For example, *mane* is a likely context of *lion* but not *animal*, even though *lion* is a hyponym of *animal*, contradicting the DIH. Rimell instead proposes measuring hyponymy using *coherence* (formalised using pointwise mutual information): the contexts of a general term minus those of a hyponym are coherent, but the reverse is not true.

Moving away from count vectors and pre-trained embeddings, there are other options. One is to build the hyponymy relation into the definition of the space. For example, Vendrov et al. (2016) use non-negative vectors, where one vector is a hyponym of another if it has a larger value in every dimension. They train a model on WordNet (Miller, 1995; Fellbaum, 1998). Building on this, Li et al. (2017) learn from both WordNet and text.

However, for a hierarchy like WordNet, there are exponentially more words lower down. This cannot be embedded in Euclidean space without words lower in the hierarchy being increasingly close together. Nickel and Kiela (2017) propose using hyperbolic space, where volume increases exponentially as we move away from any point. Țifrea et al. (2019) build on this, adapting Glove (Pennington et al., 2014) to learn hyperbolic embeddings from text. However, this approach does not generalise to non-tree hierarchies – for example, WordNet gives *bass* as a hyponym of *singer*, *voice*, *melody*, *pitch*, and *instrument*. Requiring that *bass* is represented close to all its hyperonyms also forces them close together (by the triangle inequality), which we may not want, since they are in distant parts of the hierarchy.

Alternatively, we can view hyponymy as classification, and simply use distributional vectors to provide input features (for example: Weeds et al., 2014; Rei et al., 2018). However, under this view, hyponymy is an opaque relationship, making it difficult to analyse why one vector is classified as a hyponym of another. Indeed, Levy et al. (2015) find that such classifiers mainly learn which words are common hyperonyms.

Moving away from vector representations, it can be easier to define hyponymy. Erk (2009a,b) and Gärdenfors (2014, §6.4) discuss how using regions of space provides a natural definition: P is a hy-

[14] This is also referred to as *lexical entailment*, making a link with logic (see §5.2). Other relations include antonymy, meronymy, and selectional preferences. For reasons of space, I have decided to discuss one relation in detail, rather than many relations briefly. Hyponymy could be considered basic.

ponym of Q if the region for P is contained in the region for Q. Bouraoui et al. (2017) and Vilnis et al. (2018) use this idea for knowledge base completion, and Bouraoui et al. (2020) build on this, using corpus data to identify "conceptual neighbours". In the type-driven tensorial framework (see §2), Bankova et al. (2019) and Lewis (2019) model words as normalised positive operators, with hyponymy defined in terms of subspaces (eigenspaces).

Probability distributions also allow us to define hyponymy, but it is harder than for regions, since a distribution over a smaller region has higher probability density. Vilnis and McCallum (2015) propose using KL-divergence. Athiwaratkun and Wilson (2018) propose a thresholded KL-divergence. In the type-driven tensorial framework, Balkır (2014) proposes using a quantum version of KL-divergence, which can be extended to phrases (Balkır et al., 2015; Sadrzadeh et al., 2018).

However, detecting hyponymy from corpus data remains challenging. Even in recent shared tasks (Bordea et al., 2016; Camacho-Collados et al., 2018), many systems use pattern matching, following Hearst (1992). For example, a string of the form *X such as Y* suggests that *Y* is a hyponym of *X*. In the above shared tasks, the best performing systems did not rely solely on distributional vectors, but used pattern matching as well.

Although much work remains to be done in developing learning algorithms which can detect hyponymy, I believe that a region-based approach is the most promising. Not only does it give a simple definition, but it is also motivated for other reasons, discussed elsewhere in this paper.

5 Sentence Meaning

In the previous section, I discussed meaning at the level of words. I now turn to challenges in representing meaning at the level of sentences.

5.1 Compositionality

Language is *productive* – a fluent speaker can understand a completely new sentence, as long as they know each word and each syntactic construction in the sentence. One goal for a semantic model is to be able to *derive* the meaning of a sentence from its parts, so it can generalise to new combinations. This is known as *compositionality*.[15]

For vector space models, the challenge is how to compose word vectors to construct phrase representations. If we represent both words and phrases in the same vector space, the challenge is to find a composition function that maps a pair of vectors to a new vector. In the general case, this must be sensitive to word order, since changing word order can change meaning. Mitchell and Lapata (2008, 2010) compare a variety of such functions, but find that componentwise multiplication performs best, despite being commutative, and hence insensitive to word order. The effectiveness of componentwise multiplication and addition has been replicated many times (for example: Baroni and Zamparelli, 2010; Blacoe and Lapata, 2012; Rimell et al., 2016; Czarnowska et al., 2019). However, it is unclear how to adapt it to take word order into account, and Polajnar et al. (2014) show that performance degrades with sentence length.

Alternatively, we can use a sentence space distinct from the word space. This is often done with a task-based perspective – words are combined into sentence representations, which are useful for solving some task. For example, the final state of an RNN can be seen as a representation of the whole sequence. To make the composition more linguistically informed, the network can be defined to follow a tree structure, rather than linear order (for example: Socher et al., 2010, 2012; Tai et al., 2015), or even to learn latent tree structure (for example: Dyer et al., 2016; Maillard and Clark, 2018). Alternatively, a sequence of token representations can be combined using attention, which calculates a weighted sum, as in a Transformer architecture (Vaswani et al., 2017).

Regardless of architecture, the model can be optimised either for a supervised task, such as machine translation (for example: Cho et al., 2014), or for an unsupervised objective, as in an autoencoder (for example: Hermann and Blunsom, 2013) or language model (for example: Peters et al., 2018; Devlin et al., 2019). If we take a task-based perspective, it is difficult to know if the representations will transfer to other tasks. In fact, Changpinyo et al. (2018) find that for some combinations of tasks, training on one task can be harmful for another.

As an alternative to task-based approaches, the tensorial framework mentioned in §2 also uses sentence vectors, but using tensor contraction to

[15] Kartsaklis et al. (2013) discuss how composition is often conflated with *disambiguation*, since composing ambiguous expressions often disambiguates them. Disambiguation can be seen as a kind of *contextualisation* or *context dependence*, which I discuss in §5.3. The focus in this section is on deriving semantic representations for larger expressions.

compose representations based on argument structure.[16] Polajnar et al. (2015) explore sentence spaces with dimensions defined by co-occurrences.

However, a weakness with the above approaches is that they map sentences to a finite-dimensional space. As we increase sentence length, the number of sentences with distinct meanings increases exponentially. For example, consider relative clauses: *the dog chased the cat*; *the dog chased the cat which caught the mouse*; and so on. To keep these meanings distinct, we have two options. If the meanings must be a certain distance apart, the magnitudes of sentence vectors need to increase exponentially with sentence length, so there is enough space to distinguish them.[17] Alternatively, if the meanings can be arbitrarily close, we need to record each dimension to a high precision in order to distinguish the meanings. The fine-grained structure of the space then becomes important, but small changes to model parameters (such as updates during training) would cause drastic changes to this structure. I do not know any work exploring either option. Otherwise, we are forced to view sentence vectors as lossy compression.[18] As Mooney (2014) put it: "You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector!"

Although compression can be useful for many tasks, full and detailed semantic representations also have their place. This is particularly important at a discourse level: it would be absurd to represent, as vectors of the same dimensionality, both a five-word sentence and the whole English Wikipedia. However, this leaves open the question of how we *should* represent sentence meaning. In the following section, I turn to logic as a guide.

5.2 Logic

Sentences can express complex thoughts, and build chains of reasoning. Logic formalises this, and one goal for a semantic model is to support the logical notions of *truth* (discussed in §3.2), and *entailment* (one proposition following from another).

Vectors do not have logical structure, but can still be used to provide features for a logical system, for example if entailment is framed as classification: given a *premise* and *hypothesis*, the task is to decide if the premise entails the hypothesis, contradicts it, or neither. Datasets include SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018).

However, it is difficult to analyse approaches that do not use an explicit logic. In fact, Gururangan et al. (2018) suggest that high performance may be due to annotation artifacts: only using the hypothesis, they achieve 67% on SNLI and 53% on MultiNLI, much higher than the majority class baseline (34% and 35%, respectively). Performance on such datasets may therefore overestimate the ability of neural models to perform inference.

To explicitly represent logical structure, there are a few options. One is to build a hybrid system, combining a vector space with a logic. For example, Herbelot and Vecchi (2015) aim to give logical interpretations to vectors. They consider a number of properties (such as: *is_edible*, *has_a_handle*, *made_of_wood*), and for each, they learn a mapping from vectors to values in $[0, 1]$, where 0 means the property applies to no referents, and 1 means it applies to all referents. This is an interesting way to probe what information is available in distributional vectors, but it is unclear how it could be generalised to deal with individual referents (rather than summarising them all), or to deal with complex propositions (rather than single properties).

Garrette et al. (2011) and Beltagy et al. (2016) incorporate a vector space model into a Markov Logic Network (Richardson and Domingos, 2006), a kind of probability logic. If two predicates have high distributional similarity, they add a probabilistic inference rule saying that, if one predicate is true of an entity, the other predicate is likely to also be true. This allows us to use distributional vectors in a well-defined logical model, but it assumes we can interpret similarity in terms of inference (for discussion, see: Erk, 2016). As argued in §3 above, pre-trained vectors may have already lost information, and in the long term, it would be preferable to learn logical representations directly.

Lewis and Steedman (2013) use a classical logic, and cluster predicates that are observed to hold of the same pairs of named entities – for example, *write*(*Rowling*, *Harry Potter*) and *author*(*Rowling*, *Harry Potter*). This uses corpus data directly, rather than pre-trained vectors. However, it would need to be generalised to learn from arbitrary sentences,

[16] Zanzotto et al. (2015) show how sentence similarity in this framework decomposes in terms of similarity of corresponding parts, because composition and dot products are linear.

[17] This can be formalised information-theoretically. Consider sending a message as a D-dimensional vector, through a noisy channel. If there is an upper bound K to the vector's magnitude, the channel has a finite *channel capacity*. The capacity scales as K^D, which is only polynomial in K.

[18] This conclusion has been drawn before (for example: Goodfellow et al., 2016, p. 370), but my argument makes the conditions more precise.

and not just those involving named entities.

A second option is to define a vector space with a logical interpretation. Grefenstette (2013) gives a logical interpretation to the type-driven tensorial framework (see §2), where the sentence space models truth values, and the noun space models a domain of N entities. However, Grefenstette shows that quantification would be nonlinear, so cannot be expressed using tensor contraction. Hedges and Sadrzadeh (2019) provide an alternative account which can deal with quantifiers, but at the expense of noun dimensions corresponding to *sets* of entities, so we have 2^N dimensions for N entities.

Copestake and Herbelot (2012) propose that dimensions could correspond to logical expressions being true of an entity in a situation. However, this requires generalising from an *actual* distribution (based on observed utterances) to an *ideal* distribution (based on truth of logical expressions). They do not propose a concrete algorithm, but they discuss several challenges, and suggest that grounded data might be necessary. In this vein, Kuzmenko and Herbelot (2019) use the Visual Genome dataset (Krishna et al., 2017) to learn vector representations with logically interpretable dimensions, although these vectors are not as expressive as Copestake and Herbelot's ideal distributions.

Finally, a third option is to learn logical representations instead of vectors. For example, in my own work I have represented words as truth-conditional functions that are compatible with first-order logic (Emerson and Copestake, 2017b; Emerson, 2020b). Since referents are not observed in distributional semantics, this introduces latent variables that make the model computationally expensive, although there are ways to mitigate this (Emerson, 2020a).

Despite the computational challenges, I believe the right approach is to learn a logically interpretable model, either by defining a vector space with logical structure, or by directly using logical representations. However, an important question is what kind of logic to use. I argued in §4.1 that probabilities of truth and fuzzy truth values can capture vagueness, and there are corresponding logics.

In probability logic, propositions have probabilities of being true or false, with a joint distribution for the truth values of all propositions (for an introduction, see: Adams, 1998; Demey et al., 2013). In fuzzy logic, propositions have fuzzy truth values, and classical logical operators (such as: $\wedge$, $\vee$, $\neg$) are replaced with fuzzy versions (for an introduction, see: Hájek, 1998; Cintula et al., 2017). Fuzzy operators act directly on truth values – for example, given the fuzzy truth values of p and q, we can calculate the fuzzy truth value of $p \vee q$. In contrast, in probability logic, given probabilities of truth for p and q, we cannot calculate the probability of truth for $p \vee q$, unless we know the joint distribution.

A problem with fuzzy logic, observed by Fine (1975), comes with propositions like $p \vee \neg p$. For example, suppose we have a reddish orange object, so the truth of *red* and *orange* are both below 1. Intuitively, both *red or not red* and *red or orange* should definitely be true. However, in fuzzy logic, they could have truth below 1. This makes probability logic more appealing than fuzzy logic.[19]

Furthermore, there are well-developed frameworks for probabilistic logical semantics (for example: Goodman and Lassiter, 2015; Cooper et al., 2015), which a probabilistic distributional semantics could connect to, or draw inspiration from.

5.3 Context Dependence

The flipside of compositionality is *context dependence*: the meaning of an expression often depends on the context it occurs in. For example, a *small elephant* is not a *small animal*, but a *large mouse* is – the meanings of *small* and *large* depend on the nouns they modify. One goal for a semantic model is to capture how meaning depends on context.[20]

Following Recanati (2012), we can distinguish *standing meaning*, the context-independent meaning of an expression, and *occasion meaning*, the context-dependent meaning of an expression in a particular occasion of use.[21] However, every usage occurs in *some* context, so a standing meaning must be seen as an abstraction across usages, rather than a usage in a "null" context (for discussion, see: Searle, 1980; Elman, 2009).

One approach is to treat a distributional vector as a standing meaning, and modify it to produce occasion meanings. For example, vectors could be modified according to syntactic or semantic dependencies (for example: Erk and Padó, 2008; Thater et al., 2011; Dinu et al., 2012), or even chains of

[19] Hájek et al. (1995) prove that fuzzy logic can be used to provide upper and lower bounds on probabilities in a probability logic, giving it a different motivation.

[20] Ultimately, this must include dependence on real-world context. Even the intuitive conclusion that a large mouse is a small animal depends on the implicit assumption that you and I are both humans, or at least, human-sized. From the perspective of an ant, a mouse is large animal.

[21] This terminology adapts Quine (1960).

dependencies (for example: Weir et al., 2016).

This mapping from standing vectors to occasion vectors can also be trained (for example: Czarnowska et al., 2019; Popa et al., 2019). Large language models such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) can also be interpreted like this – these models map a sequence of input embeddings to a sequence of contextualised embeddings, which can be seen as standing meanings and occasion meanings, respectively.

Alternatively, standing meanings and occasion meanings can be represented by different kinds of object. Erk and Padó (2010) represent a standing meaning as a set of vectors (each derived from a single sentence of the training corpus), and an occasion meaning is a weighted sum of these vectors.

For a probabilistic model, calculating an occasion meaning can be cast as Bayesian inference, conditioning on the context. This gives us a well-understood theoretical framework, making it easier to generalise a model to other kinds of context.

Dinu and Lapata (2010) interpret a vector as a distribution over latent senses, where each component is the probability of a sense. Given probabilities of generating context words from latent senses, we can then condition the standing distribution on the context. However this model relies on a finite sense inventory, which I argued against in §4.2.

Lui et al. (2012) and Lau et al. (2012, 2014) use LDA (Blei et al., 2003), where an occasion meaning is a distribution over context words (varying continuously as topic mixtures), and a standing meaning is a prior over such distributions.[22] A separate model is trained for each target word. Chang et al. (2014) add a generative layer, allowing them to train a single model for all target words. However, a single sense is chosen in each context, giving a finite sense inventory.

Skip-gram can be interpreted as generating context words from a target word. While we can see an embedding as a standing meaning, nothing can be seen as an occasion meaning. Bražinskas et al. (2018) add a generative layer, generating a latent vector from the target word, then generating context words from this vector. We can see a latent vector as an occasion meaning, and a word's distribution over latent vectors as a standing meaning.

Finally, in my own work, I have also calculated occasion meanings by conditioning on the context (Emerson and Copestake, 2017b), but in contrast to the above approaches, standing meanings are truth-conditional functions (binary classifiers), which I have argued for elsewhere in this paper.

[22]There are two distinct uses of a distribution here: to represent uncertainty, and to represent meaning. A sense is a topic mixture, parametrising a distribution over words; uncertainty is a Dirichlet distribution over topic mixtures.

6 Conclusion

A common thread among all of the above sections is that reaching our semantic goals requires structure beyond representing meaning as a point in space. In particular, it seems desirable to represent the meaning of a word as a region of space or as a classifier, and to work with probability logic.

However, there is a trade-off between expressiveness and learnability: the more structure we add, the more difficult it can be to work with our representations. To this end, there are promising neural architectures for working with structured data, such dependency graphs (for example: Marcheggiani and Titov, 2017) or logical propositions (for example: Rocktäschel and Riedel, 2017; Minervini et al., 2018). To mitigate computationally expensive calculations in probabilistic models, there are promising new techniques such as amortised variational inference, used in the Variational Autoencoder (Kingma and Welling, 2014; Rezende et al., 2014; Titsias and Lázaro-Gredilla, 2014).

My own recent work in this direction has been to develop the Pixie Autoencoder (Emerson, 2020a), and I look forward to seeing alternative approaches from other authors, as the field of distributional semantics continues to grow. I hope that this survey paper will help other researchers to develop the field in a way that keeps long-term goals in mind.

Acknowledgements

This paper is based on chapter 2 of my PhD thesis (Emerson, 2018). For invaluable advice on the structure and framing of that chapter, and therefore also of this paper, I want to thank my PhD supervisor Ann Copestake. I would also like to thank my PhD examiners, Katrin Erk and Paula Buttery, for feedback on that chapter, as well as Emily M. Bender, Guy Aglionby, Andrew Caines, and the NLIP reading group in Cambridge, for feedback on earlier drafts of this paper. I would like to thank ACL reviewers 1 & 3 for pointing out areas that were unclear, and reviewer 2 for their kind praise.

I am supported by a Research Fellowship at Gonville & Caius College, Cambridge.

References

Ernest W. Adams. 1998. *A Primer of Probability Logic*. Number 68 in CSLI Lecture Notes. Center for the Study of Language and Information (CSLI) Publications.

Keith Allan. 2001. *Natural language semantics*. Blackwell.

Ben Athiwaratkun and Andrew Gordon Wilson. 2017. Multimodal word distributions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 1645–1656.

Ben Athiwaratkun and Andrew Gordon Wilson. 2018. Hierarchical density order embeddings. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

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

Esma Balkır, Mehrnoosh Sadrzadeh, and Bob Coecke. 2015. Distributional sentence entailment using density matrices. In *Proceedings of the 1st International Conference on Topics in Theoretical Computer Science (TTCS)*, pages 1–22. International Federation for Information Processing (IFIP).

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

Oren Barkan. 2017. Bayesian neural word embedding. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3135–3143. Association for the Advancement of Artificial Intelligence.

Marco Baroni. 2016. Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13.

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

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 15th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1183–1193. Association for Computational Linguistics.

Lawrence W. Barsalou, Léo Dutriaux, and Christoph Scheepers. 2018. Moving beyond the distinction between concrete and abstract concepts. *Philosophical Transactions of the Royal Society B*, 373(1752):20170144.

Jon Barwise and John Perry. 1983. *Situations and Attitudes*. Massachusetts Institute of Technology (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. *Computational Linguistics*, 42(4):763–808.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Richard Bergmair. 2010. *Monte Carlo Semantics: Robust inference and logical pattern processing with Natural Language text*. Ph.D. thesis, University of Cambridge.

William Blacoe and Mirella Lapata. 2012. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, pages 546–556. Association for Computational Linguistics.

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

Alexandre Blondin-Massé, Guillaume Chicoisne, Yassine Gargouri, Stevan Harnad, Olivier Picard, and Odile Marcotte. 2008. How is meaning grounded in dictionary definitions? In *Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing*, pages 17–24. 22nd International Conference on Computational Linguistics (COLING).

Gemma Boleda, Abhijeet Gupta, and Sebastian Padó. 2017. Instances and concepts in distributional space. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Short Papers*, pages 79–85.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. SemEval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval)*, pages 1081–1091. Association for Computational Linguistics.

Zied Bouraoui, Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert. 2020. Modelling semantic categories using conceptual neighborhood. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence.

Zied Bouraoui, Shoaib Jameel, and Steven Schockaert. 2017. Inductive reasoning about ontologies using conceptual spaces. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 20th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642. Association for Computational Linguistics.

Arthur Bražinskas, Serhii Havrylov, and Ivan Titov. 2018. Embedding words as distributions with a Bayesian skip-gram model. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1775–1789. International Committee on Computational Linguistics (ICCL).

Elia Bruni, Giang Binh Tran, and Marco Baroni. 2011. Distributional semantics from text and images. In *Proceedings of GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 22–32. Association for Computational Linguistics.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research (JAIR)*, 49(2014):1–47.

Luana Bulat, Douwe Kiela, and Stephen Clark. 2016. Vision and feature norms: Improving automatic feature norm learning through cross-modal maps. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 579–588.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. SemEval-2018 task 9: Hypernym discovery. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval)*, pages 712–724. Association for Computational Linguistics.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research (JAIR)*, 63:743–788.

Ronnie Cann. 1993. *Formal semantics: an introduction*. Cambridge University Press.

João P. Carvalho, Fernando Batista, and Luisa Coheur. 2012. A critical survey on the use of fuzzy sets in speech and natural language processing. In *Proceedings of the 2012 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8.

Baobao Chang, Wenzhe Pei, and Miaohong Chen. 2014. Inducing word sense with automatically learned hidden concepts. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 355–364. International Committee on Computational Linguistics (ICCL).

Soravit Changpinyo, Hexiang Hu, and Fei Sha. 2018. Multi-task learning for sequence tagging: An empirical study. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 2965–2977.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics.

Petr Cintula, Christian G. Fermüller, and Carles Noguera. 2017. Fuzzy logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, fall edition. Metaphysics Research Lab, Stanford University.

Stephen Clark. 2015. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*, 2nd edition, chapter 16, pages 493–522. Wiley.

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

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

Ann Copestake and Aurélie Herbelot. 2012. Lexicalised compositionality. Unpublished manuscript.

Ryan Cotterell, Adam Poliak, Benjamin Van Durme, and Jason Eisner. 2017. Explaining and generalizing Skip-Gram through exponential family principal component analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Short Papers*, pages 175–181.

Paula Czarnowska, Guy Emerson, and Ann Copestake. 2019. Words are vectors, dependencies are matrices: Learning word embeddings from dependency graphs. In *Proceedings of the 13th International Conference on Computational Semantics (IWCS), Long Papers*, pages 91–102. Association for Computational Linguistics.

Lorenz Demey, Barteld Kooi, and Joshua Sack. 2013. Logic and probability. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, spring edition. Metaphysics Research Lab, Stanford University. The summer 2017 edition contains minor corrections.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Georgiana Dinu and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the 15th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1162–1172. Association for Computational Linguistics.

Georgiana Dinu, Stefan Thater, and Sören Laue. 2012. A comparison of models of word meaning in context. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 611–615.

Tiansi Dong, Chrisitan Bauckhage, Hailong Jin, Juanzi Li, Olaf Cremers, Daniel Speicher, Armin B Cremers, and Joerg Zimmermann. 2018. Imposing category trees onto word-embeddings using a geometric construction. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Michael Dummett. 1976. What is a theory of meaning? (II). In Gareth Evans and John McDowell, editors, *Truth and Meaning: Essays in Semantics*, chapter 4, pages 67–137. Clarendon Press (Oxford). Reprinted in: Dummett (1993), *Seas of Language*, chapter 2, pages 34–93.

Michael Dummett. 1978. What do I know when I know a language? Presented at the Centenary Celebrations of Stockholm University. Reprinted in: Dummett (1993), *Seas of Language*, chapter 3, pages 94–105.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A Smith. 2016. Recurrent Neural Network Grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies*, pages 199–209.

Jeffrey L. Elman. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33(4):547–582.

Guy Emerson. 2018. *Functional Distributional Semantics: Learning Linguistically Informed Representations from a Precisely Annotated Corpus*. Ph.D. thesis, University of Cambridge.

Guy Emerson. 2020a. Autoencoding pixies: Amortised variational inference with graph convolutions for Functional Distributional Semantics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Guy Emerson. 2020b. Linguists who use probabilistic models love them: Quantification in Functional Distributional Semantics. In *Proceedings of Probability and Meaning (PaM2020)*. Association for Computational Linguistics.

Guy Emerson and Ann Copestake. 2016. Functional Distributional Semantics. In *Proceedings of the 1st Workshop on Representation Learning for NLP (RepL4NLP)*, pages 40–52. Association for Computational Linguistics.

Guy Emerson and Ann Copestake. 2017a. Variational inference for logical inference. In *Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML)*, pages 53–62. Centre for Linguistic Theory and Studies in Probability (CLASP).

Guy Emerson and Ann Copestake. 2017b. Semantic composition via probabilistic model theory. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*, pages 62–77. Association for Computational Linguistics.

Katrin Erk. 2009a. Supporting inferences in semantic space: representing words as regions. In *Proceedings of the 8th International Conference on Computational Semantics (IWCS)*, pages 104–115. Association for Computational Linguistics.

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

Katrin Erk. 2010. What is word meaning, really? (And how can distributional models help us describe it?). In *Proceedings of the 2010 Workshop on Geometrical Models of Natural Language Semantics*, pages 17–26. Association for Computational Linguistics.

Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

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

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (IJCNLP), Long Papers*, pages 10–18.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

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 (EMNLP)*, pages 897–906. Association for Computational Linguistics.

Katrin Erk and Sebastian Padó. 2010. Exemplar-based models for word meaning in context. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), Short Papers*, pages 92–97.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database* [Website]. Massachusetts Institute of Technology (MIT) Press.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 91–99.

Raquel Fernández and Staffan Larsson. 2014. Vagueness and learning: A type-theoretic approach. In *Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 151–159. Association for Computational Linguistics.

Kit Fine. 1975. Vagueness, truth and logic. *Synthese*, 30(3-4):265–300.

John Rupert Firth. 1951. Modes of meaning. *Essays and Studies of the English Association*, 4:118–149. Reprinted in: Firth (1957), *Papers in Linguistics*, chapter 15, pages 190–215.

John Rupert Firth. 1957. A synopsis of linguistic theory 1930–1955. In John Rupert Firth, editor, *Studies in Linguistic Analysis*, Special volume of the Philological Society, chapter 1, pages 1–32. Blackwell.

Peter Gärdenfors. 2000. *Conceptual spaces: The geometry of thought*. Massachusetts Institute of Technology (MIT) Press.

Peter Gärdenfors. 2014. *Geometry of meaning: Semantics based on conceptual spaces*. Massachusetts Institute of Technology (MIT) Press.

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.

Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 107–114.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Massachusetts Institute of Technology (MIT) Press.

Noah D. Goodman and Daniel Lassiter. 2015. Probabilistic semantics and pragmatics: Uncertainty in language and thought. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*, 2nd edition, chapter 21, pages 655–686. Wiley.

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 (*SEM)*, pages 1–10. Association for Computational Linguistics.

Paul Grice. 1967. Logic and conversation. William James Lecture, Harvard University. Reprinted in: Peter Cole and Jerry Morgan, editors (1975), *Syntax and Semantics 3: Speech Acts*, chapter 2, pages 41–58; Donald Davidson and Gilbert Harman, editors (1975), *The Logic of Grammar*, chapter 6, pages 64–74; Grice (1989), *Studies in the Way of Words*, chapter 2, pages 22–40.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Short Papers*, pages 107–112.

Petr Hájek. 1998. *Metamathematics of Fuzzy Logic*. Number 4 in Trends in Logic. Kluwer Academic Publishers.

Petr Hájek, Lluís Godo, and Francesc Esteva. 1995. Fuzzy logic and probability. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence*, pages 237–244. Morgan Kaufmann Publishers Inc.

Patrick Hanks. 2000. Do word meanings exist? *Computers and the Humanities*, 34, Special Issue on SENSEVAL: Evaluating Word Sense Disambiguation Programs:205–215.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42:335–346.

Zellig Sabbetai Harris. 1954. Distributional structure. *Word*, 10:146–162. Reprinted in: Harris (1970), *Papers in Structural and Transformational Linguistics*, chapter 36, pages 775–794; Harris (1981), *Papers on Syntax*, chapter 1, pages 3-22.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, pages 539–545. International Committee on Computational Linguistics (ICCL).

Jules Hedges and Mehrnoosh Sadrzadeh. 2019. A generalised quantifier theory of natural language in categorical compositional distributional semantics with bialgebras. *Mathematical Structures in Computer Science*, 29(6):783–809.

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

Aurélie Herbelot and Mohan Ganesalingam. 2013. Measuring semantic content in distributional vectors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), Short Papers)*, pages 440–445.

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 20th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 22–32. Association for Computational Linguistics.

Karl Moritz Hermann and Phil Blunsom. 2013. The role of syntax in vector space models of compositional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 894–904.

Shoaib Jameel and Steven Schockaert. 2017. Modeling context words as regions: An ordinal regression approach to word embedding. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL)*, pages 123–133. Association for Computational Linguistics.

Hans Kamp and Uwe Reyle. 2013. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, volume 42 of *Studies in Linguistics and Philosophy*. Springer.

Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2013. Separating disambiguation from composition in distributional semantics. In *Proceedings of the 17th Conference on Computational Natural Language Learning (CoNLL)*, pages 114–123. Association for Computational Linguistics.

Douwe Kiela, Luana Bulat, and Stephen Clark. 2015. Grounding semantics in olfactory perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL), Short Papers*, pages 231–236.

Douwe Kiela and Stephen Clark. 2017. Learning neural audio embeddings for grounding semantics in auditory perception. *Journal of Artificial Intelligence Research (JAIR)*, 60:1003–1030.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Adam Kilgarriff. 2007. Word senses. In Eneko Agirre and Philip Edmonds, editors, *Word Sense Disambiguation: Algorithms and Applications*, chapter 2, pages 29–46. Springer.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. In *Proceedings of the NIPS 2014 Deep Learning and Representation Learning Workshop*.

Alexander Koller. 2016. Top-down and bottom-up views on success in semantics. Invited talk at the 5th Joint Conference on Lexical and Computational Semantics (*SEM).

Lingpeng Kong, Cyprien de Masson d'Autume, Wang Ling, Lei Yu, Zihang Dai, and Dani Yogatama. 2020. A mutual information maximization perspective of language representation learning. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2009. Directional distributional similarity for lexical expansion. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (IJCNLP), Short Papers*, pages 69–72.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Alexander Kuhnle, Huiyuan Xie, and Ann Copestake. 2018. How clever is the FiLM model, and how clever can it be? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 162–172.

Elizaveta Kuzmenko and Aurélie Herbelot. 2019. Distributional semantics in the real world: building word vector representations from a truth-theoretic model. In *Proceedings of the 13th International Conference on Computational Semantics (IWCS), Short Papers*, pages 16–23. Association for Computational Linguistics.

William Labov. 1973. The boundaries of words and their meanings. In Charles-James Bailey and Roger W. Shuy, editors, *New Ways of Analyzing Variation in English*, chapter 24, pages 340–371. Georgetown University Press. Reprinted in: Ralph W. Fasold, editor (1983), *Variation in the Form and Use of Language: A Sociolinguistics Reader*, chapter 3, pages 29–62, Georgetown University Press.

Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*, 25(2):335–369.

Daniel Lassiter. 2011. Vagueness as probabilistic linguistic knowledge. In Rick Nouwen, Robert van Rooij, Uli Sauerland, and Hans-Christian Schmitz, editors, *Vagueness in Communication: Revised Selected Papers from the 2009 International Workshop on Vagueness in Communication*, chapter 8, pages 127–150. Springer.

Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 259–270.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Long Papers*, pages 591–601.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 1403–1414.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal Skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 153–163.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 2177–2185.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 970–976.

David Lewis. 1970. General semantics. *Synthese*, 22:18–67.

Martha Lewis. 2019. Modelling hyponymy for DisCoCat.

Mike Lewis and Mark Steedman. 2013. Combined distributional and logical semantics. *Transactions of the Association for Computational Linguistics (TACL)*, 1:179–192.

Jiwei Li and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1722–1732.

Xiang Li, Luke Vilnis, and Andrew McCallum. 2017. Improved representation learning for predicting commonsense ontologies. In *Proceedings of the ICML 17 Workshop on Deep Structured Prediction*.

Marco Lui, Timothy Baldwin, and Diana McCarthy. 2012. Unsupervised estimation of word usage similarity. In *Proceedings of the 10th Australasian Language Technology Association Workshop (ALTA)*, pages 33–41.

Jean Maillard and Stephen Clark. 2018. Latent tree learning with differentiable parsers: Shift-reduce parsing and chart parsing. In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 13–18.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1506–1515.

Arild Matsson, Simon Dobnik, and Staffan Larsson. 2019. ImageTTR: Grounding Type Theory with Records in image classification for visual question answering. In *Proceedings of the IWCS 2019 Workshop on Computing Semantics with Types, Frames and Related Structures*, pages 55–64. Association for Computational Linguistics.

Diana McCarthy, Marianna Apidianaki, and Katrin Erk. 2016. Word sense clustering and clusterability. *Computational Linguistics*, 42(2):245–275.

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 (TACL)*, 3:103–115.

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 (ICLR), Workshop Track*.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Pasquale Minervini, Matko Bosnjak, Tim Rocktäschel, and Sebastian Riedel. 2018. Towards neural theorem proving at scale. In *Proceedings of the ICML 2018 Workshop on Neural Abstract Machines & Program Induction (NAMPI)*.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 236–244.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Raymond J. Mooney. 2014. Semantic parsing: Past, present, and future. Invited talk at the ACL 2014 Workshop on Semantic Parsing.

Gregory Murphy. 2002. *The big book of concepts*. Massachusetts Institute of Technology (MIT) Press.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 6338–6347.

Vilém Novák. 2017. Fuzzy logic in natural language processing. In *Proceedings of the 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6.

Charles K. Ogden and Ivor A. Richards. 1923. *The meaning of meaning: A study of the influence of language upon thought and of the science of symbolism*. Harcourt, Brace & World, Inc.

Diane Pecher, Inge Boot, and Saskia Van Dantzig. 2011. Abstract concepts: Sensory-motor grounding, metaphors, and beyond. In Brian Ross, editor, *The Psychology of Learning and Motivation*, volume 54, chapter 7, pages 217–248. Academic Press.

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

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Robin Piedeleu, Dimitri Kartsaklis, Bob Coecke, and Mehrnoosh Sadrzadeh. 2015. Open system categorical quantum semantics in natural language processing. In *Proceedings of the 6th Conference on Algebra and Coalgebra in Computer Science (CALCO)*, volume 35 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 270–289. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Tamara Polajnar, Laura Rimell, and Stephen Clark. 2014. Evaluation of simple distributional compositional operations on longer texts. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 4440–4443. European Language Resources Association (ELRA).

Tamara Polajnar, Laura Rimell, and Stephen Clark. 2015. An exploration of discourse-based sentence spaces for compositional distributional semantics. In *Proceedings of the EMNLP Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*, pages 1–11. Association for Computational Linguistics.

Diana Nicoleta Popa, Julien Perez, James Henderson, and Eric Gaussier. 2019. Implicit discourse relation classification with syntax-aware contextualized word representations. In *Proceedings of the 32nd International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pages 203–208.

Friedemann Pulvermüller. 2013. How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences*, 17(9):458–470.

Willard Van Orman Quine. 1960. *Word and Object*. Massachusetts Institute of Technology (MIT) Press.

Reinhard Rapp. 2004. A practical solution to the problem of automatic word sense induction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL), Interactive Poster and Demonstration Sessions*, pages 26–29.

François Recanati. 2012. Compositionality, flexibility, and context-dependence. In Wolfram Hinzen, Edouard Machery, and Markus Werning, editors, *Oxford Handbook of Compositionality*, chapter 8, pages 175–191. Oxford University Press.

Marek Rei. 2013. *Minimally supervised dependency-based methods for natural language processing*. Ph.D. thesis, University of Cambridge.

Marek Rei and Ted Briscoe. 2014. Looking for hyponyms in vector space. In *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL)*, pages 68–77. Association for Computational Linguistics.

Marek Rei, Daniela Gerz, and Ivan Vulić. 2018. Scoring lexical entailment with a supervised directional similarity network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Short Papers*, pages 638–643.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1278–1286.

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

Laura Rimell. 2014. Distributional lexical entailment by topic coherence. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Long Papers*, pages 511–519.

Laura Rimell, Jean Maillard, Tamara Polajnar, and Stephen Clark. 2016. RELPRON: A relative clause evaluation dataset for compositional distributional semantics. *Computational Linguistics*, 42(4):661–701.

Tim Rocktäschel and Sebastian Riedel. 2017. End-to-end differentiable proving. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3788–3800.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192.

Eleanor Rosch. 1978. Principles of categorization. In Eleanor Rosch and Barbara Bloom Lloyd, editors, *Cognition and categorization*, chapter 2, pages 27–48. Lawrence Erlbaum Associates. Reprinted in: Eric Margolis and Stephen Laurence, editors (1999), *Concepts: Core Readings*, chapter 8, pages 189–206.

Charles Ruhl. 1989. *On monosemy: A study in linguistic semantics*. State University of New York (SUNY) Press.

Thomas A Runkler. 2016. Generation of linguistic membership functions from word vectors. In *Proceedings of the 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 993–999.

Mehrnoosh Sadrzadeh, Dimitri Kartsaklis, and Esma Balkır. 2018. Sentence entailment in compositional distributional semantics. *Annals of Mathematics and Artificial Intelligence*, 82(4):189–218.

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 1213–1223.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

John R. Searle. 1980. The background of meaning. In John R. Searle, Ferenc Kiefer, and Manfred Bierwisch, editors, *Speech Act Theory and Pragmatics*, chapter 10, pages 221–232. D. Reidel Publishing Company.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 721–732.

Edward E. Smith and Douglas L. Medin. 1981. *Categories and Concepts*. Harvard University Press.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, pages 1201–1211. Association for Computational Linguistics.

Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2010. Learning continuous phrase representations and syntactic parsing with recursive neural networks. In *Proceedings of the NIPS 2010 Deep Learning and Unsupervised Feature Learning Workshop*.

Karen Spärck-Jones. 1964. *Synonymy and Semantic Classification*. Ph.D. thesis, University of Cambridge. Reprinted in 1986 by Edinburgh University Press.

Peter R. Sutton. 2013. *Vagueness, Communication, and Semantic Information*. Ph.D. thesis, King's College London.

Peter R. Sutton. 2015. Towards a probabilistic semantics for vague adjectives. In Henk Zeevat and Hans-Christian Schmitz, editors, *Bayesian Natural Language Semantics and Pragmatics*, chapter 10, pages 221–246. Springer.

Peter R. Sutton. 2017. Probabilistic approaches to vagueness and semantic competency. *Erkenntnis*.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from Tree-structured Long Short-Term Memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL) and the 7th International Joint Conference on Natural Language Processing (IJCNLP), Volume 1 (Long Papers)*, pages 1556–1566.

Alexandru Țifrea, Gary Bécigneul, and Octavian-Eugen Ganea. 2019. Poincaré GloVe: Hyperbolic word embeddings. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

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 (IJCNLP)*, pages 1134–1143. Asian Federation of Natural Language Processing (AFNLP).

Michalis Titsias and Miguel Lázaro-Gredilla. 2014. Doubly stochastic variational Bayes for non-conjugate inference. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1971–1979.

Ronja Utescher. 2019. Visual TTR: Modelling visual question answering in Type Theory with Records. In *Proceedings of the 13th International Conference on Computational Semantics (IWCS), Student Papers*, pages 9–14. Association for Computational Linguistics.

Kees Van Deemter. 2010. *Not exactly: In praise of vagueness*. Oxford University Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-embeddings of images and language. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*.

Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. 2018. Probabilistic embedding of knowledge graphs with box lattice measures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 263–272.

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

Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. 2014. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 2249–2259. International Committee on Computational Linguistics (ICCL).

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 1015–1021. International Committee on Computational Linguistics (ICCL).

David Weir, Julie Weeds, Jeremy Reffin, and Thomas Kober. 2016. Aligning packed dependency trees: a theory of composition for distributional semantics. *Computational Linguistics*, 42(4):727–761.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Long Papers*, pages 1112–1122.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Blackwell. Translated by Gertrude Elizabeth Margaret Anscombe. The original German text was published in 1958 under the title *Philosophische Untersuchungen*.

Lotfi A. Zadeh. 1965. Fuzzy sets. *Information and Control*, 8(3):338–353.

Lotfi A. Zadeh. 1975. The concept of a linguistic variable and its application to approximate reasoning—I. *Information Sciences*, 8(3):199–249.

Fabio Massimo Zanzotto, Lorenzo Ferrone, and Marco Baroni. 2015. When the whole is not greater than the combination of its parts: A "decompositional" look at compositional distributional semantics. *Computational Linguistics*, 41(1):165–173.

Sina Zarrieß and David Schlangen. 2017a. Is this a child, a girl or a car? Exploring the contribution of distributional similarity to learning referential word meanings. In *Proceedings of the 15th Annual Conference of the European Chapter of the Association for Computational Linguistics (EACL), Short Papers*, pages 86–91.

Sina Zarrieß and David Schlangen. 2017b. Obtaining referential word meanings from visual and distributional information: Experiments on object naming. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 243–254.