Hypernymy recognition in a neural distributional model using information gain and update semantics

ADVANCED TOPICS IN THE PHILOSOPHY OF LANGUAGE
LECTURER: ARIANNA BETTI

Mick de Neeve <mick@live.nl>
University of Amsterdam

December 23, 2022

Abstract While similarity relations can be easily identified in distributional models, hypernymy is harder because the relationship is asymmetric. Information gain has been proposed as a measure, but it is not straightforward how to use this in neural models because word co-occurences are implicit. This paper presents a method to nonetheless measure it in a (neural) Word2Vec model, but it has an additional aim: to show update semantics is a viable candidate for the underlying notion of meaning in distributional models. To this end, an epistemic dynamic semantics is implemented using the neural network, with updates between hypernymic information states. The result confirms information gain experiments to a signicant degree, and outperforms it on a small subset of transitivity triples.

1 Introduction

This paper is about distributional models of language and their relation to meaning. Distributional models exploit the idea that words with similar meanings tend to occur in similar contexts: the so-called distributional hypothesis, proposed by Harris [Har54] as *"difference of meaning correlates with difference of distribution"*. They are constructed from large corpora of texts to collect word contexts in order to construct word representations in terms of observed contexts. So while words are perhaps somewhat oddly represented in terms of their complements or absences (unless they co-occur with

themselves), distributional models are quite succesful at similarity tasks (see e.g. Lenci [Len18]). However, taking the above Harris quote at face value and noting that it characterises similarity negatively ('difference'), it is perhaps not surprising that *coffee* and *cup* tend to be given high similarity ratings while, as Hill, Reichart, and Korhonen [HRK15] point out, they do not have a great deal in common other than being used in conjunction.

Distributional models represent words as vectors in a high-dimensional space, where the axes correspond to the contexts the words in question have been observed in. Determining similarity is then a matter of calculating the angle between them, or rather the cosine (the smaller the angle, the greater the cosine), and words can then be clustered together if they have high cosine similarities (see [Len18], page 156 for an explanation, and page 153 for an illustration).

What Hill *et al.* are getting at is that (cosine) similarity as such does not really capture an identifiable semantic relationship apart from the broad notion that words are similar, but that could mean many things, such as that *coffee* and *cups* tend to feature conjunctively in a particular situation such as breakfast, while they do not share many other features (plant-based liquid vs. man-made solid as the authors point out).

One such an identifiable semantic relationship that cosine similarity fails to capture is hypernymy. For instance, *coffee* and *drink* are distributionally similar, and this could be demonstrated by computing their cosine similarity. However, the cosine similarity of the hypernym/hyponym pair (*coffee*, *drink*) is the same as for (*drink*, *coffee*), i.e. similarity is a symmetric measure while the semantic relationship is asymmetric: $\mathit{coffee}(x) \models \mathit{drink}(x)$.

This is the problem that Herbelot and Ganesalingam try to address in [HG13] by measuring the information encoded in hypernym and hyponym vectors using an information-theoretic measure (information gain) that is also asymmetric. It is the topic of the next section, after which the issue of processing hypernyms and hypernyms will be given an update semantics interpretation.

2 Background

This section introduces the main paper by Herbelot and Ganesalingam that this study is based on, where differences in semantic content between hypernym/hyponym pairs are measured in order to tell them apart, and then

describes an attempt to emulate their experiment using a neural network model that requires a different approach as such models do not store word co-occurence information explicitly. This is in preparation for the middle section about a coarse update semantics model of an agent interpreting such word pairs.

2.1 Hypernymy and information gain

In [HG13], Herbelot and Ganesalingam propose a method to tell hypernyms and hyponyms apart, based on measuring the so-called Kullback-Leibler divergence, also known as relative entropy, or the information gain of one probability distribution over another. The idea is that hyponyms are more informative than hypernyms, and that this is reflected in the information encoded in their respective word vectors. The way this is measured is broadly as follows. Word vectors are rendered as probability distributions over the observed context words as events. The authors consider two distributions per word: a prior distribution which encodes the probabilities of observing context words as such, and a conditional distribution for the probabilities of observing context words given the target word, where the target is the word encoded by the model's word vector.

The difference between a hypernym and a hyponym is that in the latter case, the conditional probability should differ more from the prior than in the former case, because more information is encoded in it. In Herbelot and Ganesalingam's words, having more information makes the vectors more distorted (page 443), which will show up as a greater Kullback-Leibler divergence D_{KL}. If for a given word, the prior is Q and the conditional P, the computation is as follows.

Equation 2.1.1.
$D_{KL}(P\|Q) = \sum_i ln(\frac{P(i)}{Q(i)}P(i))$

For a given word w, its prior Q_w and conditional P_w, Herbelot and Ganesalingam use the shorthand $KL(w) = D_{KL}(P_w\|Q_w)$. Then given a dataset with hypernym/hyponym pairs, a pair (w_1, w_2) is considered correctly recognised as (*hypernym*, *hyponym*) in case $KL(w_1) < KL(w_2)$, which reflects that the more specific term w_2 is more informative by virtue of having measurably more information encoded in its conditional.

The authors report 79.4% precision on a dataset from Baroni, Bernardi, Do, and Shan [BBDS12] from the 2012 *EACL* conference, which is also used here (and available with other resources via this paper's website at `https://`

mickdeneeve.github.io/ac/atpl). A case in point for the present study is that Herbelot and Ganesalingam use a so-called count model, meaning that they can retrieve explicit co-occurence information about contextual word observations and can hence apply this to estimate the prior as well as conditional probability distributions used in their Kullback-Leibler computations (viz. [HG13], page 443). In predict models like the neural Word2Vec model used here these co-occerences are not explicitly available, however, in the next section a method is described to approximate the required divergences between the distributions (for a description of and comparison between count and predict models, see Baroni, Dinu, and Kruszewski [BDK14]).

2.2 Information gain in Word2Vec

As noted, the model used in this study is a predict model: Word2Vec (Mikolov, Chen, Corrado, and Dean [MCCD13]) in the Skip-gram prediction direction (from words to contexts, see figure 1). It is a neural network where explicit co-occurence information that was available in [HG13]'s model is 'lost in projection' during training. In other words, the vectors' output by the model can succesfully represent words in the sense that one can compute similarities as with a count model, but the vectors are only implicitly about word co-occurence events, meaning that in order to do hypernymy recognition, a different way must be used.

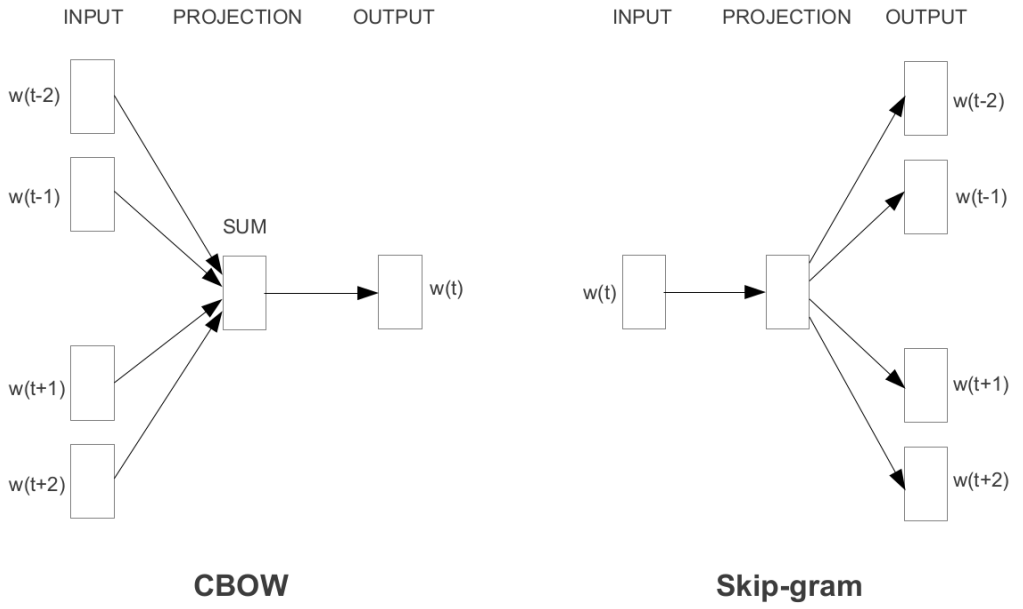

Figure 1: Word2Vec CBOW and Skip-gram architectures ([MCCD13], page 5)

A clue how to appropriately approximate the desired divergences was found in Mitra, Nalisnick, Craswell, and Caruana [MNCC16]: they point out that Word2Vec does not only have the 'ordinary' contextual word vector available, but also, for each word, the vector used to represent it during training. The paper addresses document ranking and notes that combining these vectors in measuring similarity better captures when documents are about words rather than merely mentioning them. Since the authors use the CBOW architecture they refer to these respective vectors as IN and OUT, noting that IN, i.e. the context in CBOW, is the 'normal' word vector while OUT, which represents the training word, is usually ignored.

Here, the regular word vector will be refered to as CTX ('context'), and the training word vector as WRD. The idea is that combining the two similarly results in a vector that is more strongly 'about' a word and thus more closely approximates the conditional probability P in equation 2.1.1, and that the ordinary word vector CTX can then play the role of the prior Q. Three combination methods, addition, multiplication, and maximum, were tried on a Skip-gram model with 300 dimensions and context range 4, trained on a September 2021 Wikipedia corpus of English articles (downloaded from `https://dumps.wikimedia.org/enwiki/`), and tested on Baroni *et al.*'s [BBDS12] dataset. The result is in table 1.

$P =$	$\text{CTX} + \text{WRD}$	$\text{CTX} \times \text{WRD}$	$max(\text{CTX}, \text{WRD})$
Precision:	76.4%	74.2%	72.9%

Table 1: Vector combination hypernymy experiments

With addition scoring best, the following conjecture can be formulated:

Conjecture 2.2.1.
$KL(\text{CTX}(w_1) + \text{WRD}(w_1), \text{CTX}(w_1)) < KL(\text{CTX}(w_2) + \text{WRD}(w_2), \text{CTX}(w_2))$
$\rightarrow \qquad KL(p(c_1|w_1), p(c_1)) < KL(p(c_2|w_2), p(c_2))$

This expresses the approximation: the bottom inequality is what is actually required (according to [HG13], page 442) but these probabilities are unavailable, while if the top inequality is satisfied, then w_1 is considered a hypernym of w_2.

Next, a few more models were trained to find an optimal context range (the maximum distance from target to context word), with the results in table 2.

Context range:	4	5	6	7	8
Precision:	76.4%	77.3%	78.2%	77.8%	77.0%

Table 2: Context range hypernymy experiments

With context range 6, the result of 78.2% is very close to the 79.4% achieved by Herbelot and Ganesalingam. This indicates that in (neural) predict models too, information encoded in word vectors apparently mediates semantic information. But to begin to answer the question what sort of semantics plausibly underpins this (in the next section), two observations need to be made.

First of all concerning self-information: this is a very simple hypernymy recognition measure based on the idea that more specific terms tend to occur less frequently in a corpus. Herbelot and Ganesalingam call it disappointing (page 443) that this measure outperforms their method with an 80.8% precision, but it also outperforms the method presented here at 78.6%. Though the differences are minimal, the authors' feeling that a frequency-based measure 'should not' do better is understandable since unlike information gain, it is not plausibly correlated with any notion of meaning.

Secondly, Herbelot and Ganesalingam give some misclassified examples, notably the pair (*beverage*, *beer*), which is also misclassified by the above Word2Vec method. According to the authors, while *beverage* may be a more general term, i.e. one with a larger extension, it is not general distributionally. They suspect that rather than extensions, it is intensions that distributional models provide good representations for, i.e. particular modes of speaking. *Beverage*, on this view, is misclassified because it is used in more specific contexts than *beer*.

Herbelot and Ganesalingam refer to Erk's [Erk13] claim that distributional representations stand for intensions, who in turn refers to Gärdenfors' [Gär00] suggestion that intensions are mental concepts. This appears to be in line with Herbelot and Ganesalingam as well as with Erk, since both point out that the 'classical' notion of intension as a function from possible worlds to extensions is inappropriate, as distributions such as that of *beverage* do not pick out all beverages in the world.

But distributions do seem to capture some aspects of extensions, namely

(some) properties of entities within them, e.g. *car* has more properties than its hypernym *vehicle*, inheriting properties from the latter making for a more informative vector, i.e. hypernym properties are included in hyponym properties. Baroni *et al.* [BBDS12] exploit this in a more general sense and cast vector entailment in terms of feature inclusion, where a word feature can also specifically be its co-occurences with other words.

Gärdenfors views properties as special cases of concepts ([Gär00], chapter 3), i.e. a concept applicable in a single domain. Conversely one might then say that like features, concepts are also generalisations of properties. Combining these ideas, distributions might then be viewed as mediating a mixture of extensional, intensional, and conceptual information, which may explain their success at some tasks and weakness at others. But if intensions, extensions and 'features' are all part of what a language user employs, and if these are mediated by distributions, then it is interesting to see if distributions can be used to incrementally 'construct' concepts using hyponyms and hypernyms. This is the topic of the next section, which presents an update semantics perspective as a candidate for a distributional notion of meaning.

3 Update semantics

The purpose of this section is to demonstrate the plausibility of distributional meaning, or at least aspects of it, as a type of epistemic dynamic semantics. The idea is to view hypernymic distributions from the perspective of an interpreting agent x, and to consider the process of obtaining the information encoded in the vector for a word ϕ as resulting in the modality x *knows* ϕ as being true. If ψ is a hyponym of ϕ, then since ψ is more specific and hence more informative, obtaining this information additionally should make a larger difference than in the case that ψ is a hypernym instead.

Strictly speaking, if ψ is a hypernym then learning something is ψ while already knowing it is ϕ should really make no difference at all since in that case, $\phi \models \psi$, i.e. the knowledge x holds should already accomodate ψ. However, this would be asking too much given the mixture of information distributional vectors appear to mediate (see above), so this requirement will be relaxed. In the sequel, a knowledge or information state of ignorance will be specified together with an update mechanism, resulting in the ability to measure the difference between an update with a hypernym followed by a hyponym, and a hyponym followed by a hypernym. This is then tested on the same data as that in section 2.2 (i.e. the *EACL* dataset from [BBDS12]

that is also used by [HG13]). But first some general remarks on dynamic and update semantics are in order.

3.1 The dynamic perspective

The classical notion of meaning is based on expressions in natural language having truth conditions, which boils down to the idea that objects featuring in expressions have extensions, i.e. entities in the world to which they refer. Propositions about them are true just in case these conditions are met, the classic example being due to Tarski [Tar44]: *"The sentence 'snow is white' is true if, and only if, snow is white"* (page 343). The aim here is to establish agreement with the entities (i.e. the extensions) in reality, and this agreement should be objective. But although the idea of intensions was already mooted half a century earlier by Frege in his distinction between sense and reference [Fre48], sense too, was objective for Frege as Groenenendijk and Stokhof point out in [GS00]. Sense moves the notion of meaning towards mental acts, but *"that which is grasped in such an act is essentially independent of it"* (page 2).

According to Groenendijk and Stokhof, one of the main aspects which the Fregean approach to meaning fails to account for is what they call the contextuality of utterances and their interpretation.[1] Note though that contexts in this sense is not quite the same as contexts in distributional models. A context is anything which may affect the interpretation of an utterance, such as one's knowledge of or model of the world, the time and place of the utterance, the speaker and hearer, and so on (page 6). But crucially, utterances also change contexts, in particular they may affect a hearer's model of the world (crudely put, not everyone goes out to check whether snow is indeed white). In other words, truth (or for that matter falsehoods) may also simply be asserted, and this can subsequently affect the accomodation of further information.

This last point is what matters here: the notion of context change, or information change, which is more in line with the terminology of the present paper – since change of distributional contexts as commonly understood is not the issue, but rather the change in information when distributional information is processed. And in the approach sketched by Groenendijk and Stokhof, at least when utterances are concerned which are strictly informa-

[1] The authors point out that Frege's account is restricted to what Quine called eternal sentences [Qui60].

tive, context change indeed boils down to information change. Because this process can continue incrementally, this and similar approaches are termed 'dynamic' as opposed to classic truth-conditional 'static' accounts.

The authors propose information change potential as a more suitable notion of meaning than truth conditions, since there are instances where these conditions are the same but the way information is exchanged is different, which affects interpretation. They give the following example pair: (1) *"I dropped ten marbles and found all of them, except for one. It is probably under the sofa."* (2) *"I dropped ten marbles and found only nine of them. It is probably under the sofa."* In both cases, the truth conditions are the same, but only in the first case the follow-up sentence can be succesfully accomodated. While discourses of the above complexity are well outside the scope of the present study, which only deals with simple word pairs, the message behind the example is nonetheless adopted: interpretation is information update.

The reason it is adopted is because the notion of entailment in Groenendijk, Stokhof, and Veltman [GSV96] seems to correspond well to the behaviour with respect to hypernymic information gain observed in sections 2.1 and 2.2. In [GSV96], entailment is cast as information update as follows: *"updating any information state with the premise leads to an information state in which the conclusion has to be accepted"* (page 4). This expresses that the interpreter is in a state of knowledge sufficient to accept the information provided by the conclusion (and that it *should* then be accepted).[2]

As previously noted, it will not be required here that an update with a hypernym provides no measurably new information, only that an update with a hyponym provides measurably *more* information. The notation from [GSV96]'s update definition 3.1 iii (page 12) is adopted: $s[\phi \wedge \psi] = s[\phi][\psi]$. This introduces postfix notation for sequentially updating a state s and expresses the non-commutativity of conjunction, i.e. $\mathit{cow}(x) \wedge \mathit{animal}(x)$ is not the same as $\mathit{animal}(x) \wedge \mathit{cow}(x)$ since the latter is more informative. As in *ibid* (page 11), the idea of having an initial state of ignorance will be used.

3.2 Distributive update semantics

In this section it is shown how distributional vectors can function as an interpretation model of an agent. The inspiration is ultimately Herbelot and

[2] The authors incidentally characterise static entailment as meaning inclusion, by which they essentially mean the inclusion of truth conditions, and this corresponds to property inclusion as described in section 2.2 in relation to Baroni *et al.* [BBDS12].

Ganesalingam's idea that the information in a distributional word vector mediates the semantic information of the word. The question is how this transfers from a distributional model task (hypernymy recognition) to a cognitive model of an agent that learns about objects (generally and specifically characterised), and how this affects the agent's model of the world (i.e. what Groenendijk and Stokhof call 'context'). The idea is that this model of the agent represents a mental concept, and the implication is that mental concepts are themselves distributional.

The agent's model of the world is represented by a state, which is a vector of the same dimensionality as the distributional model used. The initial state is called the *ignorant state*, as in Groenendijk, Stokhof, and Veltman. A key difference is that these authors use an eliminative approach, where as information grows, possibilities are eliminated. They mention an alternative ([GSV96], page 6) where a partial model is gradually extended, which is what is used here. The initial state then, is the uniform distribution where everything is equally likely.

Definition 3.2.1.
Let D be a dimensionality. An information state S is a vector that is a probability distribution over D, i.e. $s_1 + s_2 + \ldots + s_D \in S = 1$. If all values $s_1, \ldots, s_D = D$, then the ignorant initial state $S_0 = [\frac{1}{s_1}, \ldots, \frac{1}{s_D}]$.

As for state updates, the same combinations were tried as in section 2.2: addition, multiplication, and maximalisation. Each of them has different underlying assumptions. Addition assumes that during the information update, the entire world (or domain) is known since the result has to represent the probability mass of all that is possible. Multiplication assumes independence of the events represented in the vectors. The maximalisation function assumes that the strongest feature of a word should be adopted into new knowledge state (it is based on Gärdenfors' coincidence of events for neural networks from [Gär94]).

The results in table 3 were obtained for the Baroni *EACL* dataset, which shows addition to (again) score best. How these were computed is explained later.

Update function:	*addition*	*multiplication*	*maximalisation*
Precision:	74.1%	51.0%	65.2%

Table 3: State update hypernymy experiments

With addition selected from the attempted functions, the update of an information state can be defined, but first the concept of normalisation is needed that turns vectors into probability distributions.

Definition 3.2.2.
Let $V = [v_1, \ldots, v_n]$ be a vector. V's extent $E = v_1 + v_2 \ldots + v_n$. The normalisation of V is then given by $norm(V) = [\frac{v_1}{E}, \ldots, \frac{v_n}{E}]$.

Now an information state update will look as follows.

Definition 3.2.3.
Let S be an information state and W a word vector for a word w. Then a new state results from the update $S[w] = S[W] = norm(S + W)$. For a pair (w, u) and their vectors (W, U), the conjunction $w \wedge u$ is the sequential update $S[w \wedge u] = S[w][u] = S[W][U]$. This is equivalent to the update $S' = S[W]$ followed by $S'' = S'[U]$.

The above means, as noted, that $w \wedge u \neq u \wedge w$, an essential feature of dynamic semantics,[3] and a property capitalised on here. The main limitation of this endavour must however be made clear now, and it is one that also holds for Herbelot and Ganesalingam (and presumably many others): unlike similarity, which can be measured between random word pairs, hypernymy in their and this approach can only be determined for candidates, i.e. for pairs where it is known that one is a hypernym of the other.

Consequently, only a very confined version of entailment is possible rather than the general one that would be desirable, and it is limited to the case of such a candidate pair (it uses the shorthand of equation 2.1.1).

Definition 3.2.4.
Given a word pair (w, u) such that $w \models u \vee u \models v$, and an ignorant initial state S_0, then $u \models w$ in case the following holds: $S_1 = S_0[u]$, $S_2 = S_1[w] = S_0[u \wedge w]$, and $S'_1 = S_0[w]$, $S'_2 = S'_1[u] = S_0[w \wedge u]$, with the requirement that $KL(S_1, S_2) < KL(S'_1, S'_2)$.

The word vectors W and U implied here, this is what produced the 74.1% result on the *EACL* hypernymy dataset in table 3, and the idea behind it was already given in the previous section, but to recap: if the pair (w, u) is (*hypernym*, *hyponym*), e.g. (*vehicle*, *car*), then being told the sequence $car(x) \wedge vehicle(x)$ is less informative than $vehicle(x) \wedge car(x)$. What would

[3] Consider for instance the difference in meaning between *"John left. Mary started to cry."* and *"Mary started to cry. John left."* – the example is due to Muskens, Van Benthem, and Visser ([MvBV97], page 589).

be needed for a general entailment definition is that the sequence $car(x) \wedge vehicle(x)$ is *un*informative,[4] i.e. $vehicle(x)$ is already *accomodated*, but this is currently a bridge too far.

Finally, despite these difficulties, two concluding experiments may provide some further indication that the update semantics approach is fruitful. First of all, all hypernymy pairs succesfully recognised by information gain (in section 2.2) were run through the update semantics algorithm, which confirmed the result by 90.1%. Secondly, the *EACL* dataset contains 90 triples $v \models u \models w$, taken from hypernym/hyponym pairs (w, u) and (u, v) for which transitivity should hold. Testing whether this is the case with self-information, information gain, and update semantics produced the following result.

Measure:	*self-information*	*information gain*	*update semantics*
Precision:	37/90	41/90	45/90

Table 4: Transitivity hypernymy experiments

While this may not seem altogether brilliant, it does undo some of the painfulness of seeing the 'unsemantic' self-information measure outdo the semantically informed information gain (viz. section 2.2), and even if it is not by very much, update semantics outperforms information gain in this case.

4 Discussion

In the preceding sections, a problem with the coarseness of distributional similarity was identified, a subdomain (hypernymy) was picked that was more precise, a method was described to measure this that does justice to the asymmetry of the underlying semantics, and a cognitive model of an interpreting agent was described to try to underpin the distributional semantics encoded in a corpus with the semantic processing of an interpreting agent.

This begs the question as to the implication that the distributivity of the agent's mental state is realistic: are mental concepts as distributive as the vectors encoded in distributional models? Although distributional semantics is a more prominent research field for philosophers, linguists, and computer scientists than for psychologists, there is some evidence for this claim. For

[4] Unless the agent is in the process of learning categorisation.

instance, Miller and Charles [MC91] conjecture on the basis of similarity tasks by human resondents that word learning involves the creation of mental concept representations of contextual information.

Even though information gain experiments were contrasted with update semantics, it must be noted that the latter, too, relied on measuring state differences using information gain, but this seems fair if mental states are probabilistic.

Further research into the issue of distributive update semantics would have to look into the possibility of (total) accomodation: this could put hypernymy on a par with similarity. Similarity can be computed for random words, but hypernymy only for candidates. If a state demonstrably already accomodates an arbitrary word, then the update would be redundant since the word is entailed.

5 Concluding remarks

Similarity tasks are too coarse for determining what sort of semantics might underpin distributive models of languange; narrower tasks like hypernymy detection are more likely to be fruitful. In this paper that route was taken with a study to demonstrate how hypernymy can be identified in neural-based distributional models, and how the semantics encoded in the corpus the model was trained on might transfer to a (very limited) cognitive model of an agent processing hypernymic information.

The underlying assumption is that the mental concepts of the agent, or indeed humans, are also distributive to a particular degree, even if this is not quite the whole story.

References

[BBDS12] M. Baroni, R. Bernardi, Q.N.T. Do, and C.C. Shan. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 23–32, 2012.

[BDK14] M. Baroni, G. Dinu, and G. Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of*

the Association for Computational Linguistics, pages 238–247, 2014.

[Erk13] K. Erk. Towards a semantics for distributional representations. In *Proceedings of the 10th International Conference on Computational Semantics*. Association for Computational Linguistics, 2013.

[Fre48] G. Frege. Sense and Reference. *The Philosophical Review*, 57(3):209–230, 1948. Originally published in German (Über Sinn und Bedeutung), 1892.

[Gär94] P. Gärdenfors. How Logic Emerges from the Dynamics of Information. In J. van Eijck and A. Visser, editors, *Foundations of Computation*, pages 49–77. MIT Press, 1994.

[Gär00] P. Gärdenfors. *Conceptual Spaces*. Cambridge, MA: Bradford Books, 2000.

[GS00] J. Groenendijk and M. Stokhof. Meaning in Motion. In K. von Heusinger and U. Egli, editors, *Reference and Anaphorical Relations*, volume 72 of *Studies in Linguistics and Philosophy*. Dordrecht: Springer, 2000. Preprint (University of Amsterdam): http://dare.uva.nl/document/161517.

[GSV96] J. Groenendijk, M. Stokhof, and F. Veltman. Coreference and Modality. In S. Lapin, editor, *The Handbook of Contemporary Semantic Theory*, Blackwell Handbooks in Linguistics, pages 179–216. Oxford: Blackwell, 1996.

[Har54] Z. Harris. Distributional Structure. *Word*, 10(2-3):146–162, 1954.

[HG13] A. Herbelot and M. Ganesalingam. Measuring semantic content in distributional vectors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 440–445, 2013.

[HRK15] F. Hill, R. Reichart, and A. Korhonen. SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. *Computational Linguistics*, 41(4), 2015.

[Len18] A. Lenci. Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4:151–171, 2018.

[MC91] G.A. Miller and W.G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[MCCD13] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. Google Research, 2013. https://research.google.com/pubs/archive/41224.pdf.

[MNCC16] B. Mitra, E. Nalisnick, N. Craswell, and R. Caruana. A Dual Embedding Space Model for Document Ranking. Microsoft Research, 2016. https://www.microsoft.com/en-us/research/uploads/prod/2016/02/1602.01137v1.pdf.

[MvBV97] R. Muskens, J. van Benthem, and A. Visser. Dynamics. In *Handbook of Logic and Language*, chapter 10, pages 587–648. Elsevier Science, 1997.

[Qui60] W.V.O. Quine. *Word and Object*. MIT Press, 1960.

[Tar44] A. Tarski. The Semantic Conception of Truth and the Foundations of Semantics. *Philosophy and Phenomenological Research*, 4(3):341–376, 1944.