

Probing Taxonomic and Thematic Embeddings for Taxonomic Information

Filip Klubička and John D. Kelleher

ADAPT Centre, Technological University Dublin, Ireland

{filip.klubicka, john.kelleher}@adaptcentre.ie

Abstract

Modelling taxonomic and thematic relatedness is important for building AI with comprehensive natural language understanding. The goal of this paper is to learn more about how taxonomic information is structurally encoded in embeddings. To do this, we design a new hypernym-hyponym probing task and perform a comparative probing study of taxonomic and thematic SGNS and GloVe embeddings. Our experiments indicate that both types of embeddings encode some taxonomic information, but the amount, as well as the geometric properties of the encodings, are independently related to both the encoder architecture, as well as the embedding training data. Specifically, we find that only taxonomic embeddings carry taxonomic information in their norm, which is determined by the underlying distribution in the data.

1 Introduction

Research on probing (Ettinger et al., 2016; Shi et al., 2016; Veldhoen et al., 2016; Adi et al., 2017) has gained significant momentum in the NLP community in recent years, helping researchers explore different aspects of text encodings. While its potential for application is broad, there are still many NLP tasks the framework has not been applied to. Specifically, it seems the majority of impactful probing work focuses on analysing syntactic properties encoded in language representations, yet the rich and complex field of semantics is comparably underrepresented (Belinkov and Glass, 2019). One particular semantic problem that has not been explored at all in the context of probing is the distinction between the **taxonomic** and **thematic** dimensions of semantic relatedness (Kacmajor and Kelleher, 2019): words or concepts which belong to a common taxonomic category share properties or functions, and such relationships are commonly reflected in knowledge-engineered resources such as ontologies or taxonomies. On the other hand,

thematic relations exist by virtue of co-occurrence in a (linguistic) context where the relatedness is specifically formed between concepts performing complementary roles in a common event or theme.

This distinction informs the theoretical basis of our work, as we wish to explore the tension between taxonomic and thematic representations by examining how their information is structurally encoded. Indeed, the vast majority of pretrained language models (PTLMs) are trained solely on natural language corpora, meaning they mainly encode thematic relations. Consequently, most probing work is applied to thematic embeddings, while taxonomic embeddings remain unexplored. We thus use the probing framework to study and compare taxonomic and thematic meaning representations.

In addition, one aspect of embeddings that has not received much attention is the contribution of the vector norm to encoding linguistic information. We have recently highlighted this gap in the literature and developed an extension of the probing method called *probing with noise* (Klubička and Kelleher, 2022), which allows for relative intrinsic probe evaluations that are able to provide structural insights into embeddings and highlight the role of the vector norm in encoding linguistic information. We find taxonomic embeddings to be particularly interesting for probing the role of the norm, as we suspect that the hierarchical structure of a taxonomy is well suited to be encoded by the vector norm—given that the norm encodes the vector’s magnitude, or distance from the space’s origin, it is possible that the depth of a tree structure, such as a taxonomy, could be mapped to the vector’s distance from the origin in some way¹. Applying the *probing with noise* method to taxonomic embeddings on a taxonomic probing task could shed some light on this relationship. In order to draw

¹A hypothesis based on the finding that the squared L2 norm of BERT and ELMo can correspond to the depth of the word in a syntactic parse tree (Hewitt and Manning, 2019).

broader comparisons, we apply the same evaluation framework to taxonomic and thematic SGNS and GloVe embeddings.

2 Related Work

Hypernymy, understood as the capability to relate generic terms or classes to their specific instances, lies at the core of human cognition and plays a central role in reasoning and understanding natural language (Wellman and Gelman, 1992). Two words have a hypernymic relation if one of the words belongs to a taxonomic class that is more general than that of the other word. Hypernymy can be seen as an *IS-A* relationship, and more practically, hypernymic relations determine lexical entailment (Geffet and Dagan, 2005) and form the *IS-A* backbone of almost every ontology, semantic network and taxonomy (Yu et al., 2015). Given this, it is not surprising that modelling and identifying hypernymic relations has been pursued in NLP for over two decades (Shwartz et al., 2016).

While research on hypernym detection has been plentiful, work applying any probing framework to identify taxonomic information in embeddings is scarce, and the existing work does not probe for it directly, but rather infers taxonomic knowledge from examining higher-level tasks. For example, Ettinger (2020) identified taxonomic knowledge in BERT, but rather than using a probing classifier, BERT’s masked-LM component was used instead and its performance was examined on a range of cloze tasks. One of the relevant findings was that BERT can robustly retrieve noun hypernyms in this setting, demonstrating that BERT is strong at associating nouns with their hypernyms. Ravichander et al. (2020) build on Ettinger’s work and investigate whether probing studies shed light on BERT’s systematic knowledge, and as a case study examine hypernymy information. They devise additional cloze tasks to test for prediction consistency and demonstrate that BERT often fails to consistently make the same prediction in slightly different contexts, concluding that its ability to correctly retrieve hypernyms is not a reflection of larger systematic knowledge, but possibly an indicator of lexical memorisation (Levy et al., 2015; Santus et al., 2016; Shwartz et al., 2017).

Aside from this recent focus on BERT, little work has been done in the space of probing embeddings for hypernym information. However, work on modelling hypernymy has a long history that

stretches back before large PTLMs and includes pattern-based approaches (Hearst, 1992; Navigli and Velardi, 2010; Lenci and Benotto, 2012; Boella and Di Caro, 2013; Flati et al., 2014; Santus et al., 2014; Flati et al., 2016; Gupta et al., 2016; Pavlick and Paşca, 2017) that are based on the notion of distributional generality (Weeds et al., 2004; Clarke, 2009), as well as distributional approaches (Turney and Pantel, 2010; Baroni et al., 2012; Rei and Briscoe, 2013; Santus et al., 2014; Fu et al., 2014; Espinosa-Anke et al., 2016; Ivan Sanchez Carmona and Riedel, 2017; Nguyen et al., 2017; Pinter and Eisenstein, 2018; Bernier-Colborne and Barrière, 2018; Nickel and Kiela, 2018; Roller et al., 2018; Maldonado and Klubička, 2018; Cho et al., 2020; Mansar et al., 2021). We highlight the work of Weeds et al. (2014), who demonstrated that it is possible to predict a specific semantic relation between two words given their distributional vectors. Their work is especially relevant to ours as it shows that the nature of the relationship one is trying to establish between words informs the operation one should perform on their associated vectors, e.g. summing the vectors works well for a co-hyponym task. We consider this in §3.

In terms of evaluation benchmarks for modelling hypernymy, in most cases their design reduces them to binary classification (Baroni and Lenci, 2011; Snow et al., 2005; Boleda et al., 2017; Vyas and Carpuat, 2017), where a system has to decide whether or not a hypernymic relation holds between a given candidate pair of terms. Criticisms to this experimental setting point out that supervised systems tend to benefit from the inherent modeling of the datasets in the task, leading to lexical memorization phenomena. Some attempts to alleviate this issue involve including a graded scale for evaluating the degree of hypernymy on a given pair (Vulić et al., 2017), or reframing the task design as Hypernym Discovery (Espinosa-Anke et al., 2016). The latter addresses one of the main drawbacks of the binary evaluation criterion and resulted in the construction of a hypernym discovery benchmark covering multiple languages and knowledge domains (Camacho-Collados et al., 2018).

3 Probing Dataset Construction

Conneau et al. (2018) state that a probing task needs to ask a simple, non-ambiguous question, in order to minimise interpretability problems and confounding factors. While we acknowledge

the hypernym discovery framing as an important benchmark, and the cloze tasks used by Ettinger (2020) as an enlightening probing scenario, we suspect neither is suitable for our probing experiments, for which we require a simpler task that more directly teases out the hypernym-hyponym relationship. We thus opt to construct a new taxonomic probing task: predicting which word in a pair is the hypernym, and which is the hyponym. This dataset is directly derived from WordNet (Fellbaum, 1998) and contains all its hypernym-hyponym pairs. Thus each word pair shares only an immediate hypernym-hyponym relationship between the candidate words: a word in a pair can *only* be a hyponym or hypernym of the other.

However, in our experiments we wish to probe both taxonomic and thematic encoders. Given that we are mostly using pretrained thematic and taxonomic embeddings (see §4), their vocabulary coverage might vary dramatically. We wish to mitigate confounders by comparing like for like as much as possible, so to retain a higher integrity of interpretation when comparing models, we prune the dataset to only use the intersection of vocabularies of all the used models—we only include word pairs that have a representation for both candidate words in all the embedding models.

Note here that one of the goals of our work is to use the *probing with noise* method to learn about embeddings and the way they encode different types of information in vector space. We assert that a prediction of the relationship between a pair of words cannot be fairly done without the classifier having access to representations for both words in the pair. Yet, our probe is a classifier which can only take a single vector as input. Informed by the work of Weeds et al. (2014) we considered options such as averaging or summing the individual word vectors, but found that these were not suitable for our framing as they muddled the notion that the classifier is receiving two separate words as input. We instead concatenate the word vectors in question and pass a single concatenated vector to the classifier (similar to approaches used by Adi et al. (2017)). This approach allows us to formulate the task as a positional classification task: given a pair of words, is the first one the hypernym or the hyponym of the other? We can then assign each instance in the corpus a binary label—0 or 1—representing the class of the first word in the pair. The probe can then predict if the left half of

the vector is the hyponym (0) of the right half, or whether it is its hypernym (1).

Finally, the inherent tree structure of WordNet means that a smaller number of words will be hypernyms, while a larger number will be hyponyms. We want to avoid the probe memorising the subset of words more likely to be hypernyms, but rather to learn from information encoded in the (differences between) vectors themselves. In an attempt to achieve this, we balance out the ratio of class labels by duplicating the dataset and swapping the hypernym-hyponym positions and labels. Before duplicating, we also define a hold-out test set of 25,000 instances, so as to exclude the possibility of the same word pair appearing in both the train and test split—thus, the probe will be evaluated only on unseen instances. This duplication resulted in a final dataset of 493,494 instances, of which 50,000 comprise the test set and 443,494 comprise the training set. Here are some example instances:

0, north, direction
1, direction, north
0, hurt, upset
1, upset, hurt

4 Experimental Setup

4.1 Chosen Embeddings

In our experiments we probe taxonomic and thematic SGNS embeddings, and make an analogous comparison with taxonomic and thematic GloVe embeddings. Usually pretrained taxonomic embeddings are not as easy to come by as thematic ones, but fortunately we were able to include a set of freely available taxonomic embeddings that are based on a random walk algorithm over the WordNet taxonomy, inspired by the work of (Goikoetxea et al., 2015). In short, the approach is to generate a pseudo-corpus by crawling the WordNet structure and outputting the lexical items in the nodes visited, and then running the word embedding training on the generated pseudo-corpus. Naturally, the shape of the underlying knowledge graph affects the properties of the generated pseudo-corpus, while the types of connections that are traversed will affect the kinds of relations that are encoded in this resource. A Python implementation has been made freely available² and the embeddings

²<https://github.com/GreenParachute/wordnet-randomwalk-python>

have been shown to encode taxonomic information (Klubička et al., 2019). Ultimately we chose these embeddings as they allow us to be methodologically consistent by creating taxonomic embeddings that employ the same encoder architectures used to obtain thematic embeddings.

word2vec (SGNS) For *taxonomic SGNS* representations³ we opt for embeddings trained on the pseudo-corpus that yielded the highest Spearman correlation score on the wn-paths benchmark (introduced by Klubička et al. (2020)), i.e. the corpus with 2 million sentences, with the walk going both ways and with a 2-word minimum sentence length. The lack of a directionality constraint provides higher vocabulary coverage and a smaller proportion of rare words, while the 2-word minimum sentence length limit ensures that we only have representations for words that are part of WordNet’s taxonomic graph and have at least one hypernym-hyponym relationship, which makes them suitable for this task. For the *thematic SGNS* embeddings we use a pretrained model, and opt for the gensim⁴ word2vec implementation which was trained on a part of the Google News dataset (about 100 billion tokens) and contains 300-dimensional vectors for 3 million words and phrases⁵.

GloVe To train *taxonomic GloVe* embeddings, we use a popular Python implementation of the GloVe algorithm^{6,7} and, importantly, train it on the same 2m-both-2w/s pseudo-corpus as the above taxonomic SGNS was trained on⁸. For the *thematic GloVe* embeddings we use the original Stanford pretrained GloVe embeddings⁹, opting for the larger common crawl model, which was trained on 840 billion tokens and contains 300-dimensional embeddings for a total of 2.2 million words.

Note that when we concatenate the two word embeddings required for an instance in the train or test set, they become a 600-dimensional vector which is then passed on as input to the probe.

³<https://arrow.dit.ie/datas/12/>

⁴<https://radimrehurek.com/gensim/>

⁵word2vec-google-news-300

⁶<https://github.com/maciejkula/glove-python>

⁷We used the following training parameters: window=10, no_components=300, learning_rate=0.05, epochs=30, no_threads=2. Any other parameters are left as default.

⁸<https://arrow.dit.ie/datas/9/>

⁹<https://nlp.stanford.edu/projects/glove/>

4.2 Probing with Noise

The method is described in detail in Klubička and Kelleher (2022)¹⁰: in essence it applies targeted noise functions to embeddings that have an ablational effect and remove information encoded either in the norm or dimensions of a vector.

We remove information from the norm (abl.N) by sampling random norm values and scaling the vector dimensions to the new norm. Specifically, we sample the L2 norms uniformly from a range between the minimum and maximum L2 norm values of the respective embeddings in our dataset¹¹.

To ablate information encoded in the dimensions (abl.D), we randomly sample dimension values and then scale them to match the original norm of the vector. Specifically, we sample the random dimension values uniformly from a range between the minimum and maximum dimension values of the respective embeddings in our dataset¹². We expect this to fully remove all interpretable information encoded in the dimension values, making the norm the only information container available to the probe.

Applying both noise functions to the same vector (abl.D+N) should remove any information encoded in it, meaning the probe has no signal to learn from, a scenario equal to training on random vectors.

Even when no information is encoded in an embedding, the train set may contain class imbalance, and the probe can learn the distribution of classes. To account for this, as well as the possibility of a powerful probe detecting an empty signal (Zhang and Bowman, 2018), we need to establish informative random baselines against which we can compare the probe’s performance. We employ two such baselines: (a) we assert a random prediction (*rand.pred*) onto the test set, negating any information that a classifier could have learned, class distributions included; and (b) we train the probe on randomly generated vectors (*rand.vec*), establishing a baseline with access only to class distributions.

Importantly, while we use randomised baselines

¹⁰Code available here: <https://github.com/GreenParachute/probing-with-noise>

¹¹Thematic SGNS: [0.6854, 9.3121]

Taxonomic SGNS: [2.1666, 7.6483]

Thematic GloVe: [3.1519, 13.1196]

Taxonomic GloVe: [0.0167, 6.3104]

¹²Thematic SGNS: [-1.5547, 1.7109]

Taxonomic SGNS: [-1.8811, 1.7843]

Thematic GloVe: [-4.2095, 4.0692]

Taxonomic GloVe: [-1.3875, 1.3931]

as a sense check, we use the vanilla SGNS and GloVe word embeddings in their respective evaluations as *vanilla baselines* against which all of the introduced noise models are compared. Here, the probe has access to both dimension and norm information, as well as class distributions from the training set. However, given the lack of probing taxonomic embeddings in the literature, it is equally important to establish the vanilla baseline’s performance against the random baselines: we need to confirm that the relevant information is indeed encoded somewhere in the embeddings.

Finally, to address the degrees of randomness in the method, we train and evaluate each model 50 times and report the average score of all the runs, essentially bootstrapping over the random seeds (Wendlandt et al., 2018). Additionally, we calculate a confidence interval (CI) to make sure that the reported averages were not obtained by chance, and report it alongside the results.

4.3 Probing Classifier and Evaluation Metric

The embeddings are used as input to a Multi-Layered Perceptron (MLP) classifier, which predicts their class labels. We used the scikit-learn MLP implementation (Pedregosa et al., 2011) using the default parameters¹³. The choice of evaluation metric used to evaluate the probes is not trivial, as we want to make sure that it reliably reflects a signal captured in the embeddings, especially in an imbalanced dataset where the probe could learn the label distributions, rather than detect a true signal related to the probed phenomenon. Following our original approach (Klubička and Kelleher, 2022), we use the AUC-ROC score¹⁴, which is suited to reflecting the classifier’s performance on both positive and negative classes.

5 Experimental Results

Experimental evaluation results for taxonomic and thematic embeddings on the hypernym-hyponym probing task are presented in Tables 1 and 2. Note that all cells shaded light grey belong to the same

¹³activation='relu', solver='adam', max_iter=200, hidden_layer_sizes=100, learning_rate_init=0.001, batch_size=min(200,n_samples), early_stopping=False, weight init. $W \sim \mathcal{N}\left(0, \sqrt{6/(fan_{in} + fan_{out})}\right)$ (scikit relu default). See: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

¹⁴https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html

SGNS				
Model	THEM		TAX	
	auc	±CI	auc	±CI
rand. pred.	.5000	.0009	.4997	.0009
rand. vec.	.5001	.0012	.5001	.0011
vanilla	.9163	.0004	.9256	.0003
abl. N	.9057	.0004	.9067	.0005
abl. D	.5039	.0008	.5294	.0010
abl. D+N	.4998	.0010	.5002	.0009

Table 1: Probing results on SGNS models and baselines. Reporting average AUC-ROC scores and confidence intervals (CI) of the average of all training runs.

distribution as random baselines on a given task, as there is no statistically significant difference between the different scores; cells shaded dark grey belong to the same distribution as the vanilla baseline on a given task; and all cells that are not shaded contain a significantly different score than both the random and vanilla baselines, indicating that they belong to different distributions.

SGNS Starting with *thematic SGNS* (THEM), Table 1 shows that the random baselines perform comparably to each other, as would be expected, and their score indicates no ability to discriminate between the two classes. We can see that the vanilla representations significantly outperform the random baselines, indicating that at least some taxonomic information is encoded in the embeddings.

The norm ablation scenario (abl.N) causes a statistically significant drop in performance when compared to the vanilla baseline. In principle, this indicates that some information has been lost. If instead of the norm, we ablate the dimensions (abl.D), we see a much more dramatic performance drop compared to vanilla, indicating that much more information has been removed. Unsurprisingly, the difference in the probe’s performance when applying both noising functions (abl.D+N) compared to random baselines is not statistically significant, meaning there is no pertinent information left in these representations. Notably, once just the dimension container is ablated, its performance drops to extremely low levels and approaches random baseline performance, yet it does not quite reach it—as small as it is, the difference is statistically significant, indicating that not all information has been removed in this setting. While significant, given how minor this difference is, one might argue it does not convincingly indicate the norm’s role in

GloVe				
Model	THEM		TAX	
	auc	\pm CI	auc	\pm CI
rand. pred.	.4999	.0011	.4998	.0010
rand. vec.	.5001	.0010	.5001	.0008
vanilla	.9327	.0004	.8824	.0005
abl. N	.9110	.0004	.8435	.0008
abl. D	.5002	.0008	.6621	.0008
abl. D+N	.5000	.0011	.5006	.0011

Table 2: Probing results on GloVe models and baselines. Reporting average AUC-ROC scores and confidence intervals (CI) of the average of all training runs.

encoding taxonomic information.

However, we observe a much crisper signal in the *taxonomic SGNS* (TAX) results. The random baselines perform comparably, while the vanilla baseline significantly outperforms them, while also significantly outperforming the THEM vanilla baseline, confirming that the taxonomic embeddings encode more taxonomic information than thematic embeddings. The norm ablation scenario causes a statistically significant performance drop from vanilla, while ablating the dimension container yields a larger drop, but does not reach the random-like performance achieved when ablating both containers. Here the difference in scores between ablating just the dimensions and ablating both dimensions and norm is also significantly different from random, but notably also an order of magnitude larger than in the THEM example. This indicates that the taxonomic SGNS embeddings use the norm to encode taxonomic information more so than thematic ones.

GloVe In Table 2 we see that *thematic GloVe* (THEM) vanilla performance dramatically outperforms the baselines, but the scores drop when the norm is ablated. After ablating the dimension information, there is a substantial drop in the probe’s performance and it is immediately comparable to random baselines with no statistically significant difference. Furthermore, performance does not significantly change after also ablating the norm.

Meanwhile, the *taxonomic GloVe* embeddings tell a different story. Firstly, while vanilla embeddings outperform the random baselines, they perform much worse than THEM vanilla GloVe, indicating an inferior representation for the hypernym-hyponym prediction task, even though they were trained on WordNet random walk pseudo-corpora

(we discuss this in §6). Ablating the dimensions causes a significant drop in performance, but it is nowhere near the random performance reached when ablating both dimensions and norm. This is a really strong signal that indicates the norm encodes some hypernym-hyponym information. This echoes the findings on SGNS, showing that taxonomic embeddings tend to use the norm to encode taxonomic information more so than thematic ones.

5.1 Dataset Validation Experiments: Dimension Deletions

Our experimental design is based on the assumption that providing the probe with a concatenated vector of word embeddings would allow it to infer the asymmetric relationship between the words and use that signal to make predictions. While we have taken some steps to ensure this and mitigate lexical memorisation (see §3), there is still a concern that the models could have memorised other regularities encoded in the individual word representations and used that information to make predictions. For example, while many candidate words can indeed be both hyponyms or hypernyms, given the tree structure of the taxonomy and the distribution of edges, the frequencies at which a word takes on a hypernym or hyponym role are still skewed. It is thus more likely that any given word will be a hyponym than a hypernym, and it is possible that the embeddings implicitly encode the frequency at which a word takes on a hypernym role, versus a hyponym role.

To validate that the probe is actually learning a relationship between the candidate words, we run an additional batch of probing experiments to establish another set of baselines specific to this particular probing task. We examine the impact of two scenarios on the probe’s performance: given the same labels, a) what if the probe’s input was only one word vector, and b) what if the probe’s input was only half of each word vector in the pair?

We denote this line of enquiry as *deletion experiments*, given that in practice a) can be seen as deleting half of the concatenated vector, and b) as deleting one half each vector before concatenating. The crucial difference is that in a) the probe can only learn from one word vector without having any access to a representation of the other word, meaning it can only predict whether the candidate word is a hyponym or a hypernym by relying on the probability derived from its frequency. In b) the

SGNS				
Model	THEM		TAX	
	auc	±CI	auc	±CI
rand. pred.	.5000	.0009	.4997	.0009
rand. vec.	.5001	.0012	.5001	.0011
vanilla	.9163	.0004	.9256	.0003
del. ea. 1h	.8929	.0004	.8998*	.0005
del. ea. 2h	.8927	.0004	.9039	.0004
del. ct. 1h	.8496	.0004	.8525	.0004
del. ct. 2h	.8495	.0004	.8523	.0003

Table 3: Probing results on SGNS deletions and baselines. Reporting average AUC-ROC scores and confidence intervals (CI) of the average of all training runs.

probe has a representation for both vectors, meaning it could leverage the relationship between them, but the individual vectors are truncated, meaning that half of the dimensions are gone for each word, making this inferior to the vanilla setting¹⁵.

We ran these experiments for taxonomic and thematic SGNS and GloVe embeddings and when performing deletions assessed the impact of both halves of the vectors. All dimension deletion results are included in Tables 3 and 4, where scenario a) is denoted as *del.ct.1h/2h* (deleted 1st/2nd half of concatenated vector) and scenario b) is denoted as *del.ea.1h/2h* (deleted 1st/2nd half of each vector). When comparing the deletions of the different halves, in cases where there is a statistically significant difference between their scores, the lower of the two scores is marked with an asterisk (*).

SGNS Unsurprisingly, deleting half of the vector in either scenario causes a statistically significant drop in performance when compared to vanilla. We also observe a larger drop in both *del.ct.* settings versus the *del.ea.* settings, which confirms that predicting a word’s relationship to an “imaginary” other word is the more difficult task.

However, strikingly, the performance is also significantly above random, which indicates that the probe likely did learn some frequency distributions from the graph. It is possible that this is a reflection of the imbalance inherent to WordNet, given the large number of leaf nodes in the taxonomic graph.

Even still, the significant difference in scores between the two settings demonstrates that having access to both words, even at the cost of half the

¹⁵This choice is motivated by a desire to make this setting comparable to a) in terms of dimensionality—had we simply compared it to vanilla, it would have the advantage of having access to twice as many dimensions.

GloVe				
Model	THEM		TAX	
	auc	±CI	auc	±CI
rand. pred.	.4999	.0011	.4998	.0010
rand. vec.	.5001	.0010	.5001	.0008
vanilla	.9327	.0004	.8824	.0005
del. ea. 1h	.9120*	.0003	.8727	.0005
del. ea. 2h	.9179	.0004	.8730	.0006
del. ct. 1h	.8522	.0004	.8405	.0004
del. ct. 2h	.8522	.0004	.8406	.0004

Table 4: Probing results on GloVe deletions and baselines. Reporting average AUC-ROC scores and confidence intervals (CI) of the average of all training runs.

information in each word’s dimensions, is more informative than having a full representation of a single word, *indicating that the probe is inferring the relevant relationship between them.*

GloVe The GloVe deletion results echo the findings on SGNS in most settings. Deleting half of the vector in either scenario causes a significant performance drop, which is largely above random performance, and the drop is larger in the *del.ct.* setting versus the *del.ea.* setting. This provides further indication that, while there is an inherent imbalance in the underlying data, the probe is inferring the relevant relationship between the candidate words when given a concatenation of two word vectors. The probe benefits significantly from having access to a representation of both words, or even just two halves of each representation. Even when it is not explicitly told that it is actually getting two inputs, it is able to pick up on the fact that there is a difference between them which can be helpful in deciding on a label.

6 Discussion

There are a number of points to take away from our experimental results. Firstly, we see that both vanilla thematic embeddings encode taxonomic information and the GloVe vanilla model significantly outperforms the SGNS vanilla model. This is at least partially due to the fact that the pre-trained SGNS and GloVe thematic embeddings were trained on unrelated corpora, which differ in terms of size, topic and coverage: the corpus that GloVe was trained on is over 8 times larger than the one used to train the SGNS model, and belongs to a different, much more varied genre of text data. Thus, word representations derived from

these resources are likely very different and it is possible that due to the broader scope and much larger size of the GloVe corpus, the GloVe representations reflect more taxonomic knowledge.

However, these encoders exhibit the opposite behaviour when trained on the same WordNet random walk pseudo-corpus: expectedly, vanilla taxonomic SGNS scores improve upon its thematic version, yet vanilla taxonomic GloVe scores significantly underperform compared to thematic. While we would expect it to mirror what was observed in SGNS, taxonomic GloVe is in fact our worst-performing vanilla model. Given the significant differences in model architectures, it is possible that this unexpected behaviour is due to an interaction between the architecture and training data¹⁶. While this may play a role, we suspect that the dominant factor is rather training corpus size. The WordNet pseudo-corpus used for training taxonomic embeddings was only about 9 million tokens in size (which is sufficient to encode taxonomic relations, as shown by [Maldonado et al. \(2019\)](#)), whereas SGNS and GloVe were trained on 100 and 840 billion tokens respectively. It is not surprising that GloVe trained on a small and relatively sparse pseudo-corpus underperforms compared to training on a large natural corpus. If anything, it is encouraging that SGNS trained on a 9-million-token pseudo-corpus outperforms one trained on a 100-billion-token natural corpus.

Another important finding from our experiments is the strong evidence that *word embedding models can use the norm to encode taxonomic information, regardless of what is encoded in the vector dimensions*. We find the clearest example of this in taxonomic GloVe after ablating dimension information, where the score remains as high as ≈ 0.66 , meaning that the difference of 0.16 points is solely due to information in the norm. This is a very large difference given our understanding of the underlying mechanics, where it is well known that dimensions contain most, if not all information relevant for a task (e.g. [Durrani et al. \(2020, 2022\)](#)), and this is much more than has been demonstrated on any of the sentence-level experiments in our previous work ([Klubička and Kelleher, 2022](#)). Additionally, this is the only case where deleting half of each word vector yields a significantly higher score (≈ 0.87) than ablating the norm (≈ 0.84). This

¹⁶The interested reader might consult [Klubička \(2022, pages 121-123\)](#) for some speculation as to what that interaction might be.

suggests that more information is lost when the norm is ablated than when half of the dimensions are removed. This is a strong indicator that in this case the *norm encodes information that is not at all available in the dimensions*. Certainly, the majority of the information in an embedding is and will always be encoded in the dimensions, but it is striking how much of it is present in the norm in this case.

Generally, when it comes to dimension deletion experiments, it is expected that the performance would drop dramatically in comparison to vanilla embeddings. However, an important takeaway is that in all settings the drop is much smaller than might be expected, being quite close to vanilla performance and largely above random performance. This points to a redundancy within the dimensions themselves, seeing as either half of the vector seems to carry more than half the information required to model the task, indicating that not many dimensions are needed to encode specific linguistic features. This is consistent with the findings of ([Durrani et al., 2020](#)), who analysed individual neurons in PTLMs and found that small subsets of neurons are sufficient to predict certain linguistic tasks. Our deletion results certainly corroborate these findings, given how small the drop in the probe’s performance is when half the vector is deleted.

For additional insight into the norm, we examine the norm values. We calculate the norms of the individual hypernym and hyponym word vectors in our dataset and present the results in [Figure 1](#). The median norm value shows that the difference between hypernym and hyponym norms seems to be minor in both thematic embedding types (GloVe: 6.26 and 6.24; SGNS: 2.78 and 2.76), whereas the difference is an order of magnitude larger in both taxonomic representations (GloVe: 2.03 and 2.67; SGNS: 5.64 and 5.80). The difference is also quite large between taxonomic GloVe and SGNS, and it seems to be what is reflected in our experimental results, which show that GloVe stores the most hypernym-hyponym information in the norm.

The median norm measurements show that, on average, the norm of hypernyms is larger than the norm of hyponyms. This means that hypernyms, which are higher up in the tree, are positioned further away from the origin of the vector space than hyponyms, which are positioned lower in the tree and are closer to the origin. Notably, this is only

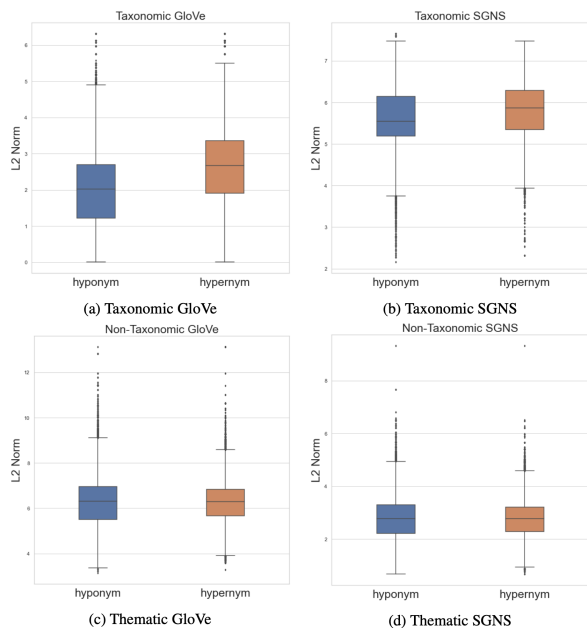


Figure 1: Box plots depicting the median values of the L2 norm in the different sets of word vectors, separate for hyponyms and hypernyms. There is a marked difference observed between hyponym and hypernym norms in taxonomic GloVe and SGNS, but not in thematic.

true in taxonomic embeddings, but not the thematic ones, indicating that in taxonomic embeddings *there is a mapping between the taxonomic hierarchy and distance from the origin*.

Finally, in spite of the fact that taxonomic GloVe (TAX) is the worst-performing vanilla model, it is interesting that its norm also encodes the most taxonomic information. We base our interpretation of this result on the following: i) in many embeddings there is a high correlation between the norm and word frequency (Goldberg, 2017), and ii) WordNet pseudo-corpora reflect hypernym-hyponym frequencies and co-occurrences. We suspect the principal signal that plays a role in the way taxonomic embeddings encode taxonomic knowledge is precisely these word co-occurrences, which GloVe is designed to capture. In turn, the norm can be seen as analogous to the hierarchical nature of taxonomic relationships and becomes the most accessible place to store this information. The thematic corpora reflect thematic co-occurrences and frequencies and hence GloVe (THEM) does not store taxonomic information in the norm, as such relations are not hierarchical in nature.

7 Conclusion

In this paper we applied the *probing with noise* method to two different types of word representations—taxonomic and thematic—each generated by two different embedding algorithms—SGNS and GloVe—on a newly-designed taxonomic probing task. The overall findings are that (a) both taxonomic and thematic static embeddings encode taxonomic information, (b) that the norm of static embedding vectors carries some taxonomic information and (c) thus the vector norm is a separate information container at the word level. (d) While in some cases there can be redundancy between the information encoded in the norm and dimensions, at other times the norm can encode information that is not at all available in the dimensions, and (e) whether the norm is utilised at all is sometimes dependant on training data, not just the encoder architecture.

We also show that in the case of SGNS, taxonomic embeddings outperform thematic ones on the task, demonstrating the usefulness of taxonomic pseudo-corpora in encoding taxonomic information. Indeed, this work serves to further emphasise the importance of the norm, showing that the taxonomic embeddings use the norm to supplement their encoding of taxonomic information. In other words, random walk corpora can improve taxonomic information in word representations, which is not always the case for natural language corpora.

Acknowledgements

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreements No. 13/RC/2106 and 13/RC/2106_P2 at the ADAPT SFI Research Centre at Technological University Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme, and is co-funded under the European Regional Development Fund.

References

- Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR, 2017*.
- Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. 2012. [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, Avignon, France. Association for Computational Linguistics.
- Marco Baroni and Alessandro Lenci. 2011. **How we BLESSEd distributional semantic evaluation**. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10, Edinburgh, UK. Association for Computational Linguistics.
- Yonatan Belinkov and James Glass. 2019. **Analysis methods in neural language processing: A survey**. *Transactions of the Association for Computational Linguistics*, 7:49–72.
- Gabriel Bernier-Colborne and Caroline Barrière. 2018. **CRIM at SemEval-2018 task 9: A hybrid approach to hypernym discovery**. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 725–731, New Orleans, Louisiana. Association for Computational Linguistics.
- Guido Boella and Luigi Di Caro. 2013. **Extracting definitions and hypernym relations relying on syntactic dependencies and support vector machines**. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 532–537, Sofia, Bulgaria. Association for Computational Linguistics.
- 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: Volume 2, Short Papers*, pages 79–85, Valencia, Spain. Association for Computational Linguistics.
- 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*, pages 712–724, New Orleans, Louisiana. Association for Computational Linguistics.
- Yejin Cho, Juan Diego Rodriguez, Yifan Gao, and Katrin Erk. 2020. **Leveraging wordnet paths for neural hypernym prediction**. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3007–3018.
- Daoud Clarke. 2009. **Context-theoretic semantics for natural language: an overview**. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 112–119, Athens, Greece. Association for Computational Linguistics.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. **What you can cram into a single $\$&!#\ast$ vector: Probing sentence embeddings for linguistic properties**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.
- Nadir Durrani, Fahim Dalvi, and Hassan Sajjad. 2022. **Linguistic correlation analysis: Discovering salient neurons in deepnlp models**. *arXiv preprint arXiv:2206.13288*.
- Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. **Analyzing individual neurons in pre-trained language models**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4865–4880, Online. Association for Computational Linguistics.
- Luis Espinosa-Anke, Jose Camacho-Collados, Claudio Delli Bovi, and Horacio Saggion. 2016. **Supervised distributional hypernym discovery via domain adaptation**. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 424–435, Austin, Texas. Association for Computational Linguistics.
- Allyson Ettinger. 2020. **What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models**. *Transactions of the Association for Computational Linguistics*, 8:34–48.
- Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. **Probing for semantic evidence of composition by means of simple classification tasks**. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. 2014. **Two is bigger (and better) than one: the Wikipedia bitaxonomy project**. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 945–955, Baltimore, Maryland. Association for Computational Linguistics.
- Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. 2016. **Multiwibi: The multilingual wikipedia bitaxonomy project**. *Artificial Intelligence*, 241:66–102.
- Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. **Learning semantic hierarchies via word embeddings**. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1199–1209, Baltimore, Maryland. 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'05)*, pages 107–114, Ann Arbor, Michigan. Association for Computational Linguistics.
- Josu Goikoetxea, Aitor Soroa, and Eneko Agirre. 2015. **Random Walks and Neural Network Language Models on Knowledge Bases**. In *Human Language Technologies: The 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1434–1439, Denver, CO.
- Yoav Goldberg. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):117.
- Amit Gupta, Francesco Piccinno, Mikhail Kozhevnikov, Marius Paşca, and Daniele Pighin. 2016. **Revisiting taxonomy induction over Wikipedia**. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2300–2309, Osaka, Japan. The COLING 2016 Organizing Committee.
- Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.
- John Hewitt and Christopher D. Manning. 2019. **A structural probe for finding syntax in word representations**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.
- V. Ivan Sanchez Carmona and Sebastian Riedel. 2017. How well can we predict hypernyms from word embeddings? a dataset-centric analysis. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017-Proceedings of Conference*, volume 2, pages 401–407. Association for Computational Linguistics.
- Magdalena Kacmajor and John D. Kelleher. 2019. **Capturing and Measuring Thematic Relatedness**. *Language Resources and Evaluation*, pages 1–38.
- Filip Klubička. 2022. *Probing with Noise: Unpicking the Warp and Weft of Taxonomic and Thematic Meaning Representations in Static and Contextual Embeddings*. Ph.D. thesis, Technological University Dublin.
- Filip Klubička, Alfredo Maldonado, and John D. Kelleher. 2019. Synthetic, yet natural: Properties of word-net random walk corpora and the impact of rare words on embedding performance. In *Proceedings of GWC2019: 10th Global WordNet Conference*.
- Filip Klubička and John D. Kelleher. 2022. Probing with noise: Unpicking the warp and weft of embeddings. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Abu Dhabi, UAE. Association for Computational Linguistics.
- Filip Klubička, Alfredo Maldonado, Abhijit Mahalunkar, and John D. Kelleher. 2020. **English word-net random walk pseudo-corpora**. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4893–4902, Marseille, France. European Language Resources Association.
- Alessandro Lenci and Giulia Benotto. 2012. **Identifying hypernyms in distributional semantic spaces**. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 75–79, Montréal, Canada. Association for Computational Linguistics.
- 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*, pages 970–976, Denver, Colorado. Association for Computational Linguistics.
- Alfredo Maldonado and Filip Klubička. 2018. **ADAPT at SemEval-2018 task 9: Skip-gram word embeddings for unsupervised hypernym discovery in specialised corpora**. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 924–927, New Orleans, Louisiana. Association for Computational Linguistics.
- Alfredo Maldonado, Filip Klubička, and John D. Kelleher. 2019. Size matters: The impact of training size in taxonomically-enriched word embeddings. *Open Computer Science*.
- Youness Mansar, Juyeon Kang, and Ismail El Maarouf. 2021. **The finsim-2 2021 shared task: Learning semantic similarities for the financial domain**. In *Companion Proceedings of the Web Conference 2021, WWW '21*, page 288–292, New York, NY, USA. Association for Computing Machinery.
- Roberto Navigli and Paola Velardi. 2010. **Learning word-class lattices for definition and hypernym extraction**. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1318–1327, Uppsala, Sweden. Association for Computational Linguistics.
- Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. **Hierarchical embeddings for hypernymy detection and directionality**. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 233–243, Copenhagen, Denmark. Association for Computational Linguistics.

- Maximillian Nickel and Douwe Kiela. 2018. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pages 3779–3788. PMLR.
- Ellie Pavlick and Marius Paşca. 2017. [Identifying 1950s American jazz musicians: Fine-grained IsA extraction via modifier composition](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2099–2109, Vancouver, Canada. Association for Computational Linguistics.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. [Scikit-learn: Machine learning in Python](#). *Journal of Machine Learning Research*, 12:2825–2830.
- Yuval Pinter and Jacob Eisenstein. 2018. [Predicting semantic relations using global graph properties](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1751, Brussels, Belgium. Association for Computational Linguistics.
- Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in bert. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102.
- Marek Rei and Ted Briscoe. 2013. [Parser lexicalisation through self-learning](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 391–400, Atlanta, Georgia. Association for Computational Linguistics.
- Stephen Roller, Douwe Kiela, and Maximilian Nickel. 2018. [Hearst patterns revisited: Automatic hypernym detection from large text corpora](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 358–363, Melbourne, Australia. Association for Computational Linguistics.
- Enrico Santus, Alessandro Lenci, Tin-Shing Chiu, Qin Lu, and Chu-Ren Huang. 2016. [Nine features in a random forest to learn taxonomical semantic relations](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 4557–4564, Portorož, Slovenia. European Language Resources Association (ELRA).
- Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. 2014. [Chasing hypernyms in vector spaces with entropy](#). In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 38–42, Gothenburg, Sweden. Association for Computational Linguistics.
- Xing Shi, Inkit Padhi, and Kevin Knight. 2016. [Does string-based neural MT learn source syntax?](#) In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.
- Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. [Improving hypernymy detection with an integrated path-based and distributional method](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2389–2398, Berlin, Germany. Association for Computational Linguistics.
- Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. 2017. [Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 65–75, Valencia, Spain. Association for Computational Linguistics.
- Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In *Advances in neural information processing systems*, pages 1297–1304.
- Peter D. Turney and Patrick Pantel. 2010. [From Frequency to Meaning: Vector Space Models of Semantics](#). *Journal of Artificial Intelligence Research*, 37:141–188.
- Sara Veldhoen, Dieuwke Hupkes, and Willem H Zuidema. 2016. Diagnostic classifiers revealing how neural networks process hierarchical structure. In *Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches (at NIPS)*.
- Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. [HyperLex: A large-scale evaluation of graded lexical entailment](#). *Computational Linguistics*, 43(4):781–835.
- Yogarshi Vyas and Marine Carpuat. 2017. [Detecting asymmetric semantic relations in context: A case-study on hypernymy detection](#). In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 33–43, Vancouver, Canada. Association for Computational Linguistics.
- Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. 2014. [Learning to distinguish hypernyms and co-hyponyms](#). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2249–2259, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.
- Julie Weeds, David Weir, and Diana McCarthy. 2004. [Characterising measures of lexical distributional similarity](#). In *COLING 2004: Proceedings of the 20th*

International Conference on Computational Linguistics, pages 1015–1021, Geneva, Switzerland. COLING.

Henry M. Wellman and Susan A. Gelman. 1992. Cognitive development: Foundational theories of core domains. *Annual review of psychology*, 43(1):337–375.

Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. [Factors influencing the surprising instability of word embeddings](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2092–2102, New Orleans, Louisiana. Association for Computational Linguistics.

Zheng Yu, Haixun Wang, Xuemin Lin, and Min Wang. 2015. Learning term embeddings for hypernymy identification. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Kelly Zhang and Samuel Bowman. 2018. [Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.