Specializing Distributional Vectors of All Words for Lexical Entailment

Aishwarya Kamath[1*], **Jonas Pfeiffer**[2*], **Edoardo M. Ponti**[3], **Goran Glavaš**[4], **Ivan Vulić**[3]
[1]Oracle Labs
[2]Ubiquitous Knowledge Processing Lab (UKP-TUDA), TU Darmstadt
[3]Language Technology Lab, TAL, University of Cambridge
[4]Data and Web Science Group, University of Mannheim
[1]aishwarya.kamath@oracle.com
[2]pfeiffer@ukp.informatik.tu-darmstadt.de
[3]{ep490,iv250}@cam.ac.uk
[4]goran@informatik.uni-mannheim.de

Abstract

Semantic specialization methods fine-tune distributional word vectors using lexical knowledge from external resources (e.g., WordNet) to accentuate a particular relation between words. However, such post-processing methods suffer from limited coverage as they affect only vectors of words *seen* in the external resources. We present the first post-processing method that specializes vectors of *all vocabulary words* – including those *unseen* in the resources – for the *asymmetric* relation of lexical entailment (LE) (i.e., hyponymy-hypernymy relation). Leveraging a partially LE-specialized distributional space, our POSTLE (i.e., *post-specialization* for LE) model learns an explicit global specialization function, allowing for specialization of vectors of unseen words, as well as word vectors from other languages via cross-lingual transfer. We capture the function as a deep feed-forward neural network: its objective re-scales vector norms to reflect the concept hierarchy while simultaneously attracting hyponymy-hypernymy pairs to better reflect semantic similarity. An extended model variant augments the basic architecture with an adversarial discriminator. We demonstrate the usefulness and versatility of POSTLE models with different input distributional spaces in different scenarios (monolingual LE and zero-shot cross-lingual LE transfer) and tasks (binary and graded LE). We report consistent gains over state-of-the-art LE-specialization methods, and successfully LE-specialize word vectors for languages without any external lexical knowledge.

1 Introduction

Word-level lexical entailment (LE), also known as the TYPE-OF or hyponymy-hypernymy relation, is a fundamental *asymmetric* lexico-semantic relation (Collins and Quillian, 1972; Beckwith et al., 1991). The set of these relations constitutes a hierarchical structure that forms the backbone of semantic networks such as WordNet (Fellbaum, 1998). Automatic reasoning about word-level LE benefits a plethora of tasks such as natural language inference (Dagan et al., 2013; Bowman et al., 2015; Williams et al., 2018), text generation (Biran and McKeown, 2013), metaphor detection (Mohler et al., 2013), and automatic taxonomy creation (Snow et al., 2006; Navigli et al., 2011; Gupta et al., 2017).

However, standard techniques for inducing word embeddings (Mikolov et al., 2013; Pennington et al., 2014; Melamud et al., 2016; Bojanowski et al., 2017; Peters et al., 2018, *inter alia*) are unable to effectively capture LE. Due to their crucial dependence on contextual information and the distributional hypothesis (Harris, 1954), they display a clear tendency towards conflating different relationships such as synonymy, antonymy, meronymy and LE and broader topical relatedness (Schwartz et al., 2015; Mrkšić et al., 2017).

To mitigate this deficiency, a standard solution is a *post-processing* step: distributional vectors are gradually refined to satisfy linguistic constraints extracted from external resources such as WordNet (Fellbaum, 1998) or BabelNet (Navigli and Ponzetto, 2012). This process, termed *retrofitting* or *semantic specialization*, is beneficial to language understanding tasks (Faruqui, 2016; Glavaš and Vulić, 2018) and is extremely versatile as it can be applied on top of any input distributional space.

Retrofitting methods, however, have a major weakness: they only *locally* update vectors of words *seen* in the external resources, while leaving vectors of all other *unseen* words unchanged, as illustrated in Figure 1. Recent work (Glavaš and Vulić, 2018; Ponti et al., 2018) has demonstrated how to specialize the *full* distributional space for the *symmetric* relation of semantic (dis)similarity. The so-called *post-specialization* model learns a

[*]Both authors contributed equally to this work.

Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pages 72–83
Florence, Italy, August 2, 2019.

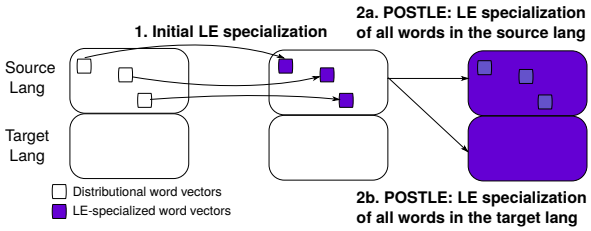


Figure 1: High-level overview of a) the POSTLE full vocabulary specialization process; and b) zero-shot cross-lingual specialization for LE. This relies on an initial shared cross-lingual word embedding space (see §2).

global and *explicit* specialization function that imitates the transformation from the distributional space to the retrofitted space, and applies it to the large subspace of unseen words' vectors.

In this work, we present POSTLE, an all-words post-specialization model for the asymmetric LE relation. This model propagates the signal on the hierarchical organization of concepts to the ones unseen in external resources, resulting in a word vector space which is fully specialized for the LE relation. Previous LE specialization methods simply integrated available LE knowledge into the input distributional space (Vulić and Mrkšić, 2018), or provided means to learn dense word embeddings of the external resource only (Nickel and Kiela, 2017, 2018; Ganea et al., 2018; Sala et al., 2018). In contrast, we show that our POSTLE method can combine distributional and external lexical knowledge and generalize over unseen concepts.

The main contribution of POSTLE is a novel global transformation function that re-scales vector norms to reflect the concept hierarchy while simultaneously attracting hyponymy-hypernymy word pairs to reflect their semantic similarity in the specialized space. We propose and evaluate two variants of this idea. The first variant learns the global function through a deep non-linear feed-forward network. The extended variant leverages the deep feed-forward net as the generator component of an adversarial model. The role of the accompanying discriminator is then to distinguish between original LE-specialized vectors (produced by any initial post-processor) from vectors produced by transforming distributional vectors with the generator.

We demonstrate that the proposed POSTLE methods yield considerable gains over state-of-the-art LE-specialization models (Nickel and Kiela, 2017; Vulić and Mrkšić, 2018), with the adversarial variant having an edge over the other. The gains are observed with different input distributional spaces in several LE-related tasks such as hypernymy detection and directionality, and graded lexical entailment. What is more, the highest gains are reported for resource-lean data scenarios where a high percentage of words in the datasets is unseen.

Finally, we show how to LE-specialize distributional spaces for target languages that lack external lexical knowledge. POSTLE can be coupled with any model for inducing cross-lingual embedding spaces (Conneau et al., 2018; Artetxe et al., 2018; Smith et al., 2017). If this model is unsupervised, the procedure effectively yields a zero-shot LE specialization transfer, and holds promise to support the construction of hierarchical semantic networks for resource-lean languages in future work.

2 Post-Specialization for LE

Our post-specialization starts with the Lexical Entailment Attract-Repel (LEAR) model (Vulić and Mrkšić, 2018), a state-of-the-art retrofitting model for LE, summarized in §2.1. While we opt for LEAR because of its strong performance and ease of use, it is important to note that our POSTLE models (§2.2 and §2.3) are not in any way bound to LEAR and can be applied on top of any LE retrofitting model.

2.1 Initial LE Specialization: LEAR

LEAR fine-tunes the vectors of words observed in a set of external linguistic constraints $C = S \cup A \cup L$, consisting of synonymy pairs S such as *(clever, intelligent)*, antonymy pairs A such as *(war, peace)*, and lexical entailment (i.e., hyponymy-hypernymy) pairs L such as *(dog, animal)*. For the L pairs, the order of words is important: we assume that the left word always refers to the hyponym.

Extending the ATTRACT-REPEL model for symmetric similarity specialization (Mrkšić et al., 2017), LEAR defines two types of objectives: 1) the ATTRACT (*Att*) objective aims to bring closer together in the vector space words that are semantically similar (i.e., synonyms and hyponym-hypernym pairs); 2) the REPEL (*Rep*) objective pushes further apart vectors of dissimilar words (i.e., antonyms). Let $\mathcal{B} = \{(\mathbf{x}_l^{(k)}, \mathbf{x}_r^{(k)})\}_{k=1}^{K}$ be the set of K word pairs for which the *Att* or *Rep* score is to be computed – these are the *positive examples*. The set of corresponding negative examples T is created by coupling each positive ATTRACT example $(\mathbf{x}_l, \mathbf{x}_r)$ with a negative example pair $(\mathbf{t}_l, \mathbf{t}_r)$, where $\mathbf{t}_l$ is the vector closest (in terms of cosine

similarity, within the batch) to $\mathbf{x}_l$ and $\mathbf{t}_r$ vector closest to $\mathbf{x}_r$. The *Att* objective for a batch of ATTRACT constraints $\mathcal{B}_A$ is then given as:

$$Att(\mathcal{B}_A, T_A) = \sum_{k=1}^{K} [\, \tau\left(\delta_{att} + \cos\left(\mathbf{x}_l^{(k)}, \mathbf{t}_l^{(k)}\right) - \cos\left(\mathbf{x}_l^{(k)}, \mathbf{x}_r^{(k)}\right)\right) + \tau\left(\delta_{att} + \cos\left(\mathbf{x}_r^{(k)}, \mathbf{t}_r^{(k)}\right) - \cos\left(\mathbf{x}_l^{(k)}, \mathbf{x}_r^{(k)}\right)\right)\,]. \quad (1)$$

$\tau(x) = max(0, x)$ is the hinge loss and δ_{att} is the similarity margin imposed between the negative and positive vector pairs. In contrast, for each positive REPEL example, the negative example $(\mathbf{t}_l, \mathbf{t}_r)$ couples the vector $\mathbf{t}_l$ that is most distant from $\mathbf{x}_l$ and $\mathbf{t}_r$, most distant from $\mathbf{x}_r$. The *Rep* objective for a batch of REPEL word pairs $\mathcal{B}_R$ is then:

$$Rep(\mathcal{B}_R, T_R) = \sum_{k=1}^{K} [\, \tau\left(\delta_{rep} + \cos\left(\mathbf{x}_l^{(k)}, \mathbf{x}_r^{(k)}\right) - \cos\left(\mathbf{x}_l^{(k)}, \mathbf{t}_l^{(k)}\right)\right) + \tau\left(\delta_{rep} + \cos\left(\mathbf{x}_l^{(k)}, \mathbf{x}_r^{(k)}\right) - \cos\left(\mathbf{x}_r^{(k)}, \mathbf{t}_r^{(k)}\right)\right)\,]. \quad (2)$$

LEAR additionally defines a regularization term in order to preserve the useful semantic information from the original distributional space. With $V(\mathcal{B})$ as the set of distinct words in a constraint batch $\mathcal{B}$, the regularization term is: $Reg(\mathcal{B}) = \lambda_{reg} \sum_{\mathbf{x} \in V(\mathcal{B})} \|\mathbf{y} - \mathbf{x}\|_2$, where $\mathbf{y}$ is the LEAR-specialization of the distributional vector $\mathbf{x}$, and λ_{reg} is the regularization factor.

Crucially, LEAR forces specialized vectors to reflect the asymmetry of the LE relation with an asymmetric distance-based objective. The goal is to preserve the cosine distances in the specialized space while steering vectors of more general concepts (those found higher in the concept hierarchy) to take larger norms.[1] Vulić and Mrkšić (2018) test several asymmetric objectives, and we adopt the one reported to be the most robust:

$$LE(\mathcal{B}_L) = \sum_{k=1}^{K} \frac{\|\mathbf{x}_l^{(k)}| - \|\mathbf{x}_r^{(k)}\|}{\|\mathbf{x}_l^{(k)}\| + \|\mathbf{x}_r^{(k)}\|}. \quad (3)$$

$\mathcal{B}_L$ denotes a batch of LE constraints. The full LEAR objective is then defined as:

$$J = Att(\mathcal{B}_S, T_S) + Rep(\mathcal{B}_A, T_A) + Att(\mathcal{B}_L, T_L) + LE(\mathcal{B}_L) + Reg(\mathcal{B}_S, \mathcal{B}_A, \mathcal{B}_L) \quad (4)$$

[1] E.g., while *dog* and *animal* should be close in the LE-specialized space in terms of cosine distance, the vector norm of *animal* should be larger than that of *dog*.

In summary, LEAR pulls words from synonymy and LE pairs closer together ($Att(\mathcal{B}_S, T_S)$ and $Att(\mathcal{B}_L, T_L)$), while simultaneously pushing vectors of antonyms further apart ($Rep(\mathcal{B}_A, T_A)$) and enforcing asymmetric distances for hyponymy-hypernymy pairs ($LE(\mathcal{B}_L)$).

2.2 Post-Specialization Model

The retrofitting model (LEAR) specializes vectors only for a subset of the full vocabulary: the words it has *seen* in the external lexical resource. Such resources are still fairly incomplete, even for major languages (e.g., WordNet for English), and fail to cover a large portion of the distributional vocabulary (referred to as *unseen* words). The transformation of the seen subspace, however, provides evidence on the desired effects of LE-specialization. We seek a post-specialization procedure for LE (termed POSTLE) that propagates this useful signal to the subspace of unseen words and LE-specializes the entire distributional space (see Figure 1).

Let $\mathbf{X}_s$ be the subset of the distributional space containing vectors of words seen in lexical constraints and let $\mathbf{Y}_s$ denote LE-specialized vectors of those words produced by the initial LE specialization model. For seen words, we pair their original distributional vectors $\mathbf{x}_s \in \mathbf{X}_s$ with corresponding LEAR-specialized vectors $\mathbf{y}_s$: post-specialization then directly uses pairs $(\mathbf{x}_s, \mathbf{y}_s)$ as training instances for learning a global specialization function, which is then applied to LE-specialize the remainder of the distributional space, i.e., the specialization function learned from $(\mathbf{X}_s, \mathbf{Y}_s)$ is applied to the subspace of unseen words' vectors $\mathbf{X}_u$.

Let $G(\mathbf{x}_i; \theta_G) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ (with d as the dimensionality of the vector space) be the specialization function we are trying to learn using pairs of distributional and LEAR-specialized vectors as training instances. We first instantiate the post-specialization model $G(\mathbf{x}_i; \theta_G)$ as a deep fully-connected feed-forward network (DFFN) with H hidden layers and m units per layer. The mapping of the j-th hidden layer is given as:

$$\mathbf{x}^{(j)} = activ\left(\mathbf{x}^{(j-1)}\mathbf{W}^j + \mathbf{b}^{(j)}\right). \quad (5)$$

$activ$ refers to a non-linear activation function,[2]

[2] As discussed by Vulić et al. (2018); Ponti et al. (2018), non-linear transformations yield better results: linear transformations cannot fully capture the subtle fine-tuning done by the retrofitting process, guided by millions of pairwise constraints. We also verify that linear transformations yield poorer performance, but we do not report these results for brevity.

$\mathbf{x}^{(j-1)}$ is the output of the previous layer ($\mathbf{x}^{(0)}$ is the input distributional vector), and $(\mathbf{W}^{(j)}, \mathbf{b}^{(j)})$, $j \in \{1, \ldots, H\}$ are the model's parameters θ_G.

The aim is to obtain predictions $G(\mathbf{x}_s; \theta_G)$ that are as close as possible to the corresponding LEAR-specializations $\mathbf{y}_s$. For symmetric similarity-based post-specialization prior work relied on cosine distance to measure discrepancy between the predicted and expected specialization (Vulić et al., 2018). Since we are specializing vectors for the asymmetric LE relation, the predicted vector $G(\mathbf{x}_s; \theta_G)$ has to match $\mathbf{y}_s$ not only in direction (as captured by cosine distance) but also in size (i.e., the vector norm). Therefore, the POSTLE objective augments cosine distance $dcos$ with the absolute difference of $G(\mathbf{x}_s; \theta_G)$ and $\mathbf{y}_s$ norms:[3]

$$\mathcal{L}_S = dcos\left(G(\mathbf{x}_s; \theta_G), \mathbf{y}_s\right) + \delta_n \big| \|G(\mathbf{x}_s; \theta_G)\| - \|\mathbf{y}_s\| \big|. \quad (6)$$

The hyperparameter δ_n determines the contribution of the norm difference to the overall loss.

2.3 Adversarial LE Post-Specialization

We next extend the DFFN post-specialization model with an adversarial architecture (ADV), following Ponti et al. (2018) who demonstrated its usefulness for similarity-based specialization. The intuition behind the adversarial extension is as follows: the specialization function $G(\mathbf{x}_s; \theta_G)$ should not only produce vectors that have high cosine similarity and similar norms with corresponding LEAR-specialized vectors $\mathbf{y}_s$, but should also ensure that these vectors seem "natural", that is, as if they were indeed sampled from $\mathbf{Y}_s$. We can force the post-specialized vectors $G(\mathbf{x}_s; \theta_G)$ to be legitimate samples from the $\mathbf{Y}_s$ distribution by introducing an adversary that learns to discriminate whether a given vector has been generated by the specialization function or directly sampled from $\mathbf{Y}_s$. Such adversaries prevent the generation of unrealistic outputs, as demonstrated in computer vision (Pathak et al., 2016; Ledig et al., 2017; Odena et al., 2017).

The DFFN function $G(\mathbf{x}; \theta_G)$ from §2.2 can be seen as the generator component. We couple the generator with the discriminator $D(\mathbf{x}; \theta_D)$, also instantiated as a DFFN. The discriminator performs binary classification: presented with a word vector, it predicts whether it has been produced by G or sampled from the LEAR-specialized subspace $\mathbf{Y}_s$. On the other hand, the generator tries to produce vectors which the discriminator would misclassify as sampled from $\mathbf{Y}_s$. The discriminator's loss is defined via negative log-likelihood over two sets of inputs; generator produced vectors $G(\mathbf{x}_s; \theta_G)$ and LEAR specializations $\mathbf{y}_s$:

$$\mathcal{L}_D = -\sum_{s=1}^{N} \log P(\text{spec} = 0 | G(\mathbf{x}_s; \theta_G); \theta_D) - \sum_{s=1}^{M} \log P(\text{spec} = 1 | \mathbf{y}_s; \theta_D) \quad (7)$$

Besides minimizing the similarity-based loss $\mathcal{L}_S$, the generator has the additional task of confusing the discriminator: it thus perceives the discriminator's correct predictions as its additional loss $\mathcal{L}_G$:

$$\mathcal{L}_G = -\sum_{s=1}^{N} \log P(\text{spec} = 1 | G(\mathbf{x}_s; \theta_G); \theta_D) - \sum_{s=1}^{M} \log P(\text{spec} = 0 | \mathbf{y}_s; \theta_D) \quad (8)$$

We learn G's and D's parameters with stochastic gradient descent – to reduce the co-variance shift and make training more robust, each batch contains examples of the same class (either only predicted vectors or only LEAR vectors). Moreover, for each update step of $\mathcal{L}_G$ we alternate between s_D update steps for $\mathcal{L}_D$ and s_S update steps for $\mathcal{L}_S$.

2.4 Cross-Lingual LE Specialization Transfer

The POSTLE models enable LE specialization of vectors of words unseen in lexical constraints. Conceptually, this also allows for a LE-specialization of a distributional space of another language (possibly without any external constraints), provided a shared bilingual distributional word vector space. To this end, we can resort to any of the methods for inducing shared cross-lingual vector spaces (Ruder et al., 2018). What is more, most recent methods successfully learn the shared space without any bilingual signal (Conneau et al., 2018; Artetxe et al., 2018; Chen and Cardie, 2018; Hoshen and Wolf, 2018).

Let $\mathbf{X}_t$ be the distributional space of some target language for which we have no external lexical constraints and let $P(\mathbf{x}; \theta_P) : \mathbb{R}^{d_t} \mapsto \mathbb{R}^{d_s}$ be the (linear) function projecting vectors $\mathbf{x}_t \in \mathbf{X}_t$ to the distributional space $\mathbf{X}_{ds}$ of the source language with available lexical constraints for which

[3] Simply minimizing Euclidean distance also aligns vectors in terms of both direction and size. However, we consistently obtained better results by the objective function from Eq. (6).

we trained the post-specialization model. We then simply obtain the LE-specialized space $\mathbf{Y}_t$ of the target language by composing the projection P with the post-specialization G (see Figure 1):

$$\mathbf{Y}_t = G(P(\mathbf{X}_t; \theta_P); \theta_G) \quad (9)$$

In §4.3 we report on language transfer experiments with three different linear projection models P in order to verify the robustness of the cross-lingual LE-specialization transfer.[4]

3 Experimental Setup

Distributional Vectors. To test the robustness of the POSTLE approach, we experiment with two pre-trained English word vector spaces: (1) vectors trained by Levy and Goldberg (2014) on the Polyglot Wikipedia (Al-Rfou et al., 2013) using Skip-Gram with Negative Sampling (SGNS-BOW2) (Mikolov et al., 2013) and (2) GLOVE embeddings trained on the Common Crawl (Pennington et al., 2014). In the cross-lingual transfer experiments (§4.3), we use English, Spanish, and French FASTTEXT embeddings trained on respective Wikipedias (Bojanowski et al., 2017).

Linguistic Constraints. We use the same set of constraints as LEAR in prior work (Vulić and Mrkšić, 2018): synonymy and antonymy constraints from (Zhang et al., 2014; Ono et al., 2015) are extracted from WordNet and Roget's Thesaurus (Kipfer, 2009). As in other work on LE specialization (Nguyen et al., 2017; Nickel and Kiela, 2017), asymmetric LE constraints are extracted from WordNet, and we collect both direct and indirect LE pairs (i.e., *(parrot, bird)*, *(bird, animal)*, and *(parrot, animal)* are in the LE set) In total, we work with 1,023,082 pairs of synonyms, 380,873 pairs of antonyms, and 1,545,630 LE pairs.

Training Configurations. For LEAR, we adopt the hyperparameter setting reported in the original paper: $\delta_{att} = 0.6$, $\delta_{rep} = 0$, $\lambda_{reg} = 10^{-9}$. For POSTLE, we fine-tune the hyperparameters via random search on the validation set: 1) DFFN uses $H = 4$ hidden layers, each with $1,536$ units and Swish as the activation function (Ramachandran et al., 2018); 2) ADV relies on $H = 4$ hidden layers, each with $m = 2,048$ units and Leaky ReLU (slope 0.2) (Maas et al., 2014) for the generator. The discriminator uses $H = 2$ layers with $1,024$ units and Leaky ReLU. For each update based on the generator loss ($\mathcal{L}_G$), we perform $s_S = 3$ updates based on the similarity loss ($\mathcal{L}_S$) and $s_D = 5$ updates based on the discriminator loss ($\mathcal{L}_D$). The value for the norm difference contribution in $\mathcal{L}_S$ is set to to $\delta_n = 0.1$ (see Eq. (6)) for both POSTLE variants. We train POSTLE models using SGD with the batch size 32, the initial learning rate 0.1, and a decay rate of 0.98 applied every 1M examples.

Asymmetric LE Distance. The distance that measures the strength of the LE relation in the specialized space reflects both the cosine distance between the vectors as well as the asymmetric difference between their norms (Vulić and Mrkšić, 2018):

$$I_{LE}(\mathbf{x}, \mathbf{y}) = dcos(\mathbf{x}, \mathbf{y}) + \frac{\|\mathbf{x}\| - \|\mathbf{y}\|}{\|\mathbf{x}\| + \|\mathbf{y}\|} \quad (10)$$

LE-specialized vectors of general concepts obtain larger norms than vectors of specific concepts. True LE pairs should display both a small cosine distance and a negative norm difference. Therefore, in different LE tasks we can rank the candidate pairs in the ascending order of their asymmetric LE distance I_{LE}. The LE distances are trivially transformed into binary LE predictions, using a binarization threshold t: if $I_{LE}(\mathbf{x}, \mathbf{y}) < t$, we predict that LE holds between words x and y with vectors $\mathbf{x}$ and $\mathbf{y}$.

4 Evaluation and Results

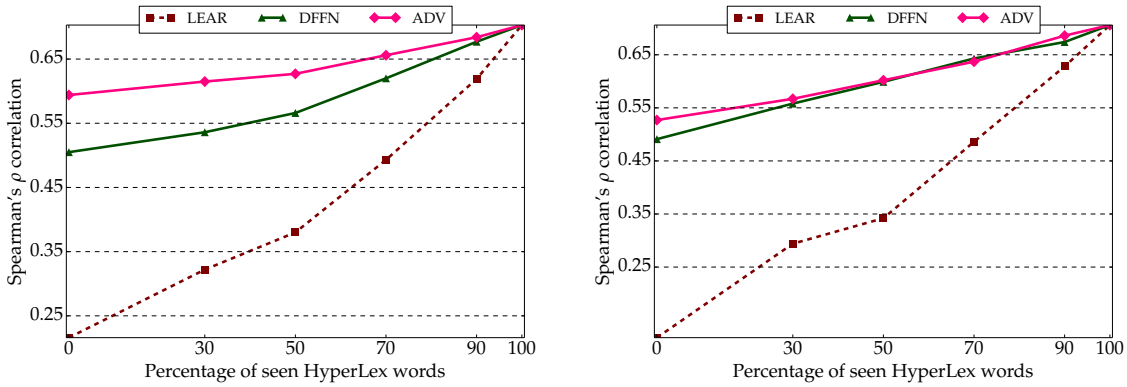
We extensively evaluate the proposed POSTLE models on two fundamental LE tasks: 1) predicting graded LE and 2) LE detection (and directionality), in monolingual and cross-lingual transfer settings.

4.1 Predicting Graded LE

The asymmetric distance I_{LE} can be directly used to make fine-grained graded assertions about the hierarchical relationships between concepts. Following previous work (Nickel and Kiela, 2017; Vulić and Mrkšić, 2018), we evaluate graded LE on the standard HyperLex dataset (Vulić et al., 2017).[5] HyperLex contains 2,616 word pairs (2,163 noun pairs, the rest are verb pairs) rated by humans by

[4] We experiment with unsupervised and weakly supervised models for inducing cross-lingual embedding spaces. However, we stress that the POSTLE specialization transfer is equally applicable on top of any method for inducing cross-lingual word vectors, some of which may require more bilingual supervision (Upadhyay et al., 2016; Ruder et al., 2018).

[5] Graded LE is a phenomenon deeply rooted in cognitive science and linguistics: it captures the notions of *concept prototypicality* (Rosch, 1973; Medin et al., 1984) and *category vagueness* (Kamp and Partee, 1995; Hampton, 2007). We refer the reader to the original paper for a more detailed discussion.



(a) SGNS-BOW2 (b) GLOVE

Figure 2: Spearman's ρ correlation scores for two input distributional spaces on the noun portion of HyperLex (2,163 concept pairs) conditioned on the number of test words covered (i.e., *seen*) in the external lexical resource. Similar patterns are observed on the full HyperLex dataset. Two other baseline models report the following scores on the noun portion of HyperLex in the 100% setting: 0.512 (Nickel and Kiela, 2017); 0.540 (Nguyen et al., 2017).

estimating on a [0, 6] scale the *degree* to which the first concept is *a type of* the second concept.

Results and Discussion. We evaluate the performance of LE specialization models in a deliberately controlled setup: we (randomly) select a percentage of HyperLex words (0%, 30%, 50%, 70%, 90% and 100%) which are allowed to be *seen* in the external constraints, and discard the constraints containing other HyperLex words, making them effectively unseen by the initial LEAR model. In the 0% setting all constraints containing any of the HyperLex words have been removed, whereas in the 100% setting, all available constraints are used. The scores are summarized in Figure 2.

The 0% setting is especially indicative of POSTLE performance: we notice large gains in performance without seeing a single word from HyperLex in the external resource. This result verifies that the POSTLE models can generalize well to words unseen in the resources. Intuitively, the gap between POSTLE and LEAR is reduced in the settings where LEAR "sees" more words. In the 100% setting we report the same scores for LEAR and POSTLE: this is an artefact of the HyperLex dataset construction as all HyperLex word pairs were sampled from WordNet (i.e., the coverage of test words is 100%). Another finding is that in the resource-leaner 0% and 30% settings POSTLE outperforms two other baselines (Nguyen et al., 2017; Nickel and Kiela, 2017), despite the fact that the two baselines have "seen" all HyperLex words. The results further indicate that POSTLE yields gains on top of different initial distributional spaces. As expected, the scores are higher with the more sophisticated ADV variant.

4.2 LE Detection

Detection and Directionality Tasks. We now evaluate POSTLE models on three binary classification datasets commonly used for evaluating LE models (Roller et al., 2014; Shwartz et al., 2017; Nguyen et al., 2017), compiled into an integrated benchmark by Kiela et al. (2015).[6]

The first task, LE directionality, is evaluated on 1,337 true LE pairs (DBLESS) extracted from BLESS (Baroni and Lenci, 2011). The task tests the models' ability to predict which word in the LE pair is the hypernym. This is simply achieved by taking the word with a larger word vector norm as the hypernym. The second task, LE detection, is evaluated on the WBLESS dataset (Weeds et al., 2014), comprising 1,668 word pairs standing in one of several lexical relations (LE, meronymy-holonymy, co-hyponymy, reverse LE, and no relation). The models have to distinguish true LE pairs from pairs that stand in other relations (including the reverse LE). We score all pairs using the I_{LE} distance. Following Nguyen et al. (2017), we find the threshold t via cross-validation.[7] Finally, we evaluate LE detection and directionality simultaneously on BIBLESS, a relabeled variant of WBLESS. The task is to detect true LE pairs (including the reverse LE pairs), and also to determine the relation directionality. We again use I_{LE} to detect LE pairs, and then compare the vector norms to select the hypernym.

For all three tasks, we consider two evaluation

[6] http://www.cl.cam.ac.uk/∼dk427/generality.html

[7] In each of the 1,000 iterations, 2% of the pairs are sampled for threshold tuning, and the remaining 98% are used for testing. The reported numbers are therefore averaged scores.

	Setup: FULL						Setup: DISJOINT					
	DBLESS		WBLESS		BIBLESS		DBLESS		WBLESS		BIBLESS	
	SG	GL	SG	GL	SG	GL	SG	GL	SG	GL	SG	GL
LEAR (Vulić et al., 2018)	.957	.955	.905	.910	.872	.875	.528	.531	.555	.529	.381	.389
POSTLE DFFN	.957	.955	.905	.910	.872	.875	.898	.825	.754	.746	.696	.677
POSTLE ADV	.957	.955	.905	.910	.872	.875	**.942**	**.888**	**.832**	**.766**	**.757**	**.690**

Table 1: Accuracy of POSTLE models on *BLESS datasets, for two different sets of English distributional vectors: Skip-Gram (SG) and GloVe (GL). LEAR reports highest scores on *BLESS datasets in the literature.

	Target: SPANISH			Target: FRENCH		
Random	.498			.515		
Distributional	.362			.387		
	Ar	Co	Sm	Ar	Co	Sm
POSTLE DFFN	**.798**	.740	.728	.688	.735	.742
POSTLE ADV	.768	**.790**	**.782**	**.746**	**.770**	**.786**

Table 2: Average precision (AP) of POSTLE models in cross-lingual transfer. Results are shown for both POSTLE models (DFFN and ADV), two target languages (Spanish and French) and three methods for inducing bilingual vector spaces: *Ar* (Artetxe et al., 2018), *Co* (Conneau et al., 2018), and *Sm* (Smith et al., 2017).

settings: 1) in the FULL setting we use all available lexical constraints (see §3) for the initial LEAR specialization; 2) in the DISJOINT setting, we remove all constraints that contain any of the test words, making all test words effectively unseen by LEAR.

Results and Discussion. The accuracy scores on *BLESS test sets are provided in Table 1.[8] Our POSTLE models display exactly the same performance as LEAR in the FULL setting: this is simply because *all* words found in *BLESS datasets are covered by the lexical constraints, and POSTLE does not generalize the initial LEAR transformation to unseen test words. In the DISJOINT setting, however, LEAR is left "blind" as it has not seen a single test word in the constraints: it leaves distributional vectors of *BLESS test words identical. In this setting, LEAR performance is equivalent to the original distributional space. In contrast, learning to generalize the LE specialization function from LEAR-specializations of other words, POSTLE models are able to successfully LE-specialize vectors of test *BLESS words. As in the graded LE, the adversarial POSTLE architecture outperforms the simpler DFFN model.

[8]We have evaluated the prediction performance also in terms of F_1 and, in the ranking formulation, in terms of *average precision* (AP) and observed the same trends in results.

4.3 Cross-Lingual Transfer

Finally, we evaluate cross-lingual transfer of LE specialization. We train POSTLE models using distributional (FASTTEXT) English (EN) vectors as input. Afterwards, we apply those models to the distributional vector spaces of two other languages, French (FR) and Spanish (ES), after mapping them into the same space as English as described in §2.4.

We experiment with several methods to induce cross-lingual word embeddings: 1) MUSE, an adversarial unsupervised model fine-tuned with the closed-form Procustes solution (Conneau et al., 2018); 2) an unsupervised self-learning algorithm that iteratively bootstraps new bilingual seeds, initialized according to structural similarities of the monolingual spaces (Artetxe et al., 2018); 3) an orthogonal linear mapping with inverse softmax, supervised by 5K bilingual seeds (Smith et al., 2017).

We test POSTLE-specialized Spanish and French word vectors on WN-Hy-ES and WN-Hy-FR, two equally sized datasets (148K word pairs) created by Glavaš and Ponzetto (2017) using the ES WordNet (Gonzalez-Agirre et al., 2012) and the FR WordNet (Sagot and Fišer, 2008). We perform a ranking evaluation: the aim is to rank LE pairs above pairs standing in other relations (meronyms, synonyms, antonyms, and reverse LE). We rank word pairs in the ascending order based on I_{LE}, see Eq. (10).

Results and Discussion. The average precision (AP) ranking scores achieved via cross-lingual transfer of POSTLE are shown in Table 2. We report AP scores using three methods for cross-lingual word embedding induction, and compare their performance to two baselines: 1) random word pair scoring, and 2) the original (FASTTEXT) vectors.

The results uncover the inability of distributional vectors to capture LE – they yield lower performance than the random baseline, which strongly emphasizes the need for the LE-specialization. The transferred POSTLE yields an immense improve-

ment over the distributional baselines (up to +0.428, i.e. +118%). Again, the adversarial architecture surpasses DFFN across the board, with the single exception of EN-ES transfer coupled with Artetxe et al. (2018)'s cross-lingual model. Furthermore, transfers with unsupervised (Ar, Co) and supervised bilingual mapping (Sm) yield comparable performance. This implies that a robust LE-specialization of distributional vectors for languages with no lexico-semantic resources is possible even without any bilingual signal or translation effort.

5 Related Work

Vector Space Specialization. In general, lexical specialization models fall into two categories: 1) joint optimization models and 2) post-processing or retrofitting models. Joint models integrate external constraints directly into the distributional objective of embedding algorithms such as Skip-Gram and CBOW (Mikolov et al., 2013), or Canonical Correlation Analysis (Dhillon et al., 2015). They either modify the prior or regularization of the objective (Yu and Dredze, 2014; Xu et al., 2014; Kiela et al., 2015) or augment it with factors reflecting external lexical knowledge (Liu et al., 2015; Ono et al., 2015; Osborne et al., 2016; Nguyen et al., 2017). Each joint model is tightly coupled to a specific distributional objective: any change to the underlying distributional model requires a modification of the whole joint model and expensive retraining.

In contrast, retrofitting models (Faruqui et al., 2015; Rothe and Schütze, 2015; Wieting et al., 2015; Jauhar et al., 2015; Nguyen et al., 2016; Mrkšić et al., 2016; Mrkšić et al., 2017; Vulić and Mrkšić, 2018) use external constraints to post-hoc fine-tune distributional spaces. Effectively, this makes them applicable to any input distributional space, but they modify only vectors of words seen in the external resource. Nonetheless, retrofitting models tend to outperform joint models in the context of both similarity-based (Mrkšić et al., 2016) and LE specialization (Vulić and Mrkšić, 2018).

The recent post-specialization paradigm has been so far applied only to the symmetric semantic similarity relation. Vulić et al. (2018) generalize over the retrofitting ATTRACT-REPEL (AR) model (Mrkšić et al., 2017) by learning a global similarity-focused specialization function implemented as a DFFN. Ponti et al. (2018) further propose an adversarial post-specialization architecture. In this work, we show that post-specialization represents a viable methodology for specializing all distributional word vectors for the LE relation as well.

Modeling Lexical Entailment. Extensive research effort in lexical semantics has been dedicated to automatic detection of the fundamental taxonomic LE relation. Early approaches (Weeds et al., 2004; Clarke, 2009; Kotlerman et al., 2010; Lenci and Benotto, 2012, *inter alia*) detected LE word pairs by means of asymmetric direction-aware mechanisms such as distributional inclusion hypothesis (Geffet and Dagan, 2005), and concept informativeness and generality (Herbelot and Ganesalingam, 2013; Santus et al., 2014; Shwartz et al., 2017), but were surpassed by more recent methods that leverage word embeddings.

Embedding-based methods either 1) induce LE-oriented vector spaces using text (Vilnis and McCallum, 2015; Yu et al., 2015; Vendrov et al., 2016; Henderson and Popa, 2016; Nguyen et al., 2017; Chang et al., 2018; Vulić and Mrkšić, 2018) and/or external hierarchies (Nickel and Kiela, 2017, 2018; Sala et al., 2018) or 2) use distributional vectors as features for supervised LE detection models (Baroni et al., 2012; Tuan et al., 2016; Shwartz et al., 2016; Glavaš and Ponzetto, 2017; Rei et al., 2018). Our POSTLE method belongs to the first group.

Vulić and Mrkšić (2018) proposed LEAR, a retrofitting LE model which displays performance gains on a spectrum of graded and ungraded LE evaluations compared to joint specialization models (Nguyen et al., 2017). However, LEAR still specializes only the vectors of words seen in external resources. The same limitation holds for a family of recent models that embed concept hierarchies (i.e., trees or directed acyclic graphs) in hyperbolic spaces (Nickel and Kiela, 2017; Chamberlain et al., 2017; Nickel and Kiela, 2018; Sala et al., 2018; Ganea et al., 2018). Although hyperbolic spaces are arguably more suitable for embedding hierarchies than the Euclidean space, the "Euclidean-based" LEAR has been proven to outperform the hyperbolic embedding of the WordNet hierarchy across a range of LE tasks (Vulić and Mrkšić, 2018).

The proposed POSTLE framework 1) mitigates the limited coverage issue of retrofitting LE-specialization models, and 2) removes the problem of dependence on distributional objective in joint models. Unlike retrofitting models, POSTLE LE-specializes vectors of *all* vocabulary words, and unlike joint models, it is computationally inexpensive and applicable to any distributional vector space.

6 Conclusion

We have presented POSTLE, a novel neural post-specialization framework that specializes distributional vectors of all words – including the ones unseen in external lexical resources – to accentuate the hierarchical asymmetric lexical entailment (LE or hyponymy-hypernymy) relation. The benefits of our two all-words POSTLE model variants have been shown across a range of graded and binary LE detection tasks on standard benchmarks. What is more, we have indicated the usefulness of the POSTLE paradigm for zero-shot cross-lingual LE specialization of word vectors in target languages, even without having any external lexical knowledge in the target. In future work, we will experiment with more sophisticated neural architectures, other resource-lean languages, and bootstrapping approaches to LE specialization. Code and POSTLE-specialized vectors are available at: [https://github.com/ashkamath/POSTLE].

Acknowledgments

EMP and IV are supported by the ERC Consolidator Grant LEXICAL (648909). The authors would like to thank the anonymous reviewers for their helpful suggestions.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of CoNLL*, pages 183–192.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL*, pages 789–798.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of EACL*, pages 23–32.

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop*, pages 1–10.

Richard Beckwith, Christiane Fellbaum, Derek Gross, and George A. Miller. 1991. WordNet: A lexical database organized on psycholinguistic principles. *Lexical acquisition: Exploiting on-line resources to build a lexicon*, pages 211–231.

Or Biran and Kathleen McKeown. 2013. Classifying taxonomic relations between pairs of Wikipedia articles. In *Proceedings of IJCNLP*, pages 788–794.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pages 632–642.

Benjamin Paul Chamberlain, James Clough, and Marc Peter Deisenroth. 2017. Neural embeddings of graphs in hyperbolic space. *CoRR*, abs/1705.10359.

Haw-Shiuan Chang, Ziyun Wang, Luke Vilnis, and Andrew McCallum. 2018. Distributional inclusion vector embedding for unsupervised hypernymy detection. In *Proceedings of NAACL-HLT*, pages 485–495.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of EMNLP*, pages 261–270.

Daoud Clarke. 2009. Context-theoretic semantics for natural language: An overview. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics (GEMS)*, pages 112–119.

Allan M. Collins and Ross M. Quillian. 1972. Experiments on semantic memory and language comprehension. *Cognition in Learning and Memory*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of ICLR (Conference Track)*.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Paramveer S. Dhillon, Dean P. Foster, and Lyle H. Ungar. 2015. Eigenwords: Spectral word embeddings. *Journal of Machine Learning Research*, 16:3035–3078.

Manaal Faruqui. 2016. *Diverse Context for Learning Word Representations*. Ph.D. thesis, Carnegie Mellon University.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL-HLT*, pages 1606–1615.

Christiane Fellbaum. 1998. *WordNet*. MIT Press.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. 2018. Hyperbolic entailment cones for learning hierarchical embeddings. In *Proceedings of ICML*, pages 1632–1641.

Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of ACL*, pages 107–114.

Goran Glavaš and Ivan Vulić. 2018. Explicit retrofitting of distributional word vectors. In *Proceedings of ACL*, pages 34–45.

Goran Glavaš and Simone Paolo Ponzetto. 2017. Dual tensor model for detecting asymmetric lexico-semantic relations. In *Proceedings of EMNLP*, pages 1758–1768.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0. In *LREC*, pages 2525–2529.

Amit Gupta, Rémi Lebret, Hamza Harkous, and Karl Aberer. 2017. Taxonomy induction using hypernym subsequences. In *Proceedings of CIKM*, pages 1329–1338.

James A. Hampton. 2007. Typicality, graded membership, and vagueness. *Cognitive Science*, 31(3):355–384.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

James Henderson and Diana Popa. 2016. A vector space for distributional semantics for entailment. In *Proceedings of ACL*, pages 2052–2062.

Aurélie Herbelot and Mohan Ganesalingam. 2013. Measuring semantic content in distributional vectors. In *Proceedings of ACL*, pages 440–445.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of EMNLP*, pages 469–478.

Sujay Kumar Jauhar, Chris Dyer, and Eduard H. Hovy. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *Proceedings of NAACL*, pages 683–693.

Hans Kamp and Barbara Partee. 1995. Prototype theory and compositionality. *Cognition*, 57(2):129–191.

Douwe Kiela, Laura Rimell, Ivan Vulić, and Stephen Clark. 2015. Exploiting image generality for lexical entailment detection. In *Proceedings of ACL*, pages 119–124.

Barbara Ann Kipfer. 2009. *Roget's 21st Century Thesaurus (3rd Edition)*. Philip Lief Group.

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

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of CVPR*, pages 4681–4690.

Alessandro Lenci and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of *SEM*, pages 75–79.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL*, pages 302–308.

Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. 2015. Learning semantic word embeddings based on ordinal knowledge constraints. In *Proceedings of ACL*, pages 1501–1511.

Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. 2014. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of ICML*.

Douglas L. Medin, Mark W. Altom, and Timothy D. Murphy. 1984. Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology*, 10(3):333–352.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. Context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of CoNLL*, pages 51–61.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119.

Michael Mohler, David Bracewell, Marc Tomlinson, and David Hinote. 2013. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 27–35.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Maria Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of NAACL-HLT*, pages 142–148.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the ACL*, 5:309–324.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli, Paola Velardi, and Stefano Faralli. 2011. A graph-based algorithm for inducing lexical taxonomies from scratch. In *Proceedings of IJCAI*, pages 1872–1877.

Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Hierarchical embeddings for hypernymy detection and directionality. In *Proceedings of EMNLP*, pages 233–243.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. In *Proceedings of ACL*, pages 454–459.

Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *Proceedings of NIPS*, pages 6341–6350.

Maximilian Nickel and Douwe Kiela. 2018. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In *Proceedings of ICML*, pages 3776–3785.

Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of ICML*, pages 2642–2651.

Masataka Ono, Makoto Miwa, and Yutaka Sasaki. 2015. Word Embedding-based Antonym Detection using Thesauri and Distributional Information. In *Proceedings of NAACL*, pages 984–989.

Dominique Osborne, Shashi Narayan, and Shay Cohen. 2016. Encoding prior knowledge with eigenword embeddings. *Transactions of the ACL*, 4:417–430.

Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of CVPR*, pages 2536–2544.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of 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 NAACL-HLT*, pages 2227–2237.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of EMNLP*, pages 282–293.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2018. Searching for activation functions. In *Proceedings of ICML*.

Marek Rei, Daniela Gerz, and Ivan Vulić. 2018. Scoring lexical entailment with a supervised directional similarity network. In *Proceedings of ACL*, pages 638–643.

Stephen Roller, Katrin Erk, and Gemma Boleda. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of COLING*, pages 1025–1036.

Eleanor H. Rosch. 1973. Natural categories. *Cognitive Psychology*, 4(3):328–350.

Sascha Rothe and Hinrich Schütze. 2015. AutoExtend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of ACL*, pages 1793–1803.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2018. A survey of cross-lingual embedding models. *Journal of Artificial Intelligence Research*.

Benoît Sagot and Darja Fišer. 2008. Building a free french wordnet from multilingual resources. In *Proccedings of the OntoLex Workshop*.

Frederic Sala, Christopher De Sa, Albert Gu, and Christopher Ré. 2018. Representation tradeoffs for hyperbolic embeddings. In *Proceedings of ICML*, pages 4457–4466.

Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. 2014. Chasing hypernyms in vector spaces with entropy. In *Proceedings of EACL*, pages 38–42.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of CoNLL*, pages 258–267.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of ACL*, pages 2389–2398.

Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. 2017. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Proceedings of EACL*, pages 65–75.

Samuel L. Smith, David H.P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR (Conference Track)*.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of ACL*, pages 801–808.

Luu Anh Tuan, Yi Tay, Siu Cheung Hui, and See Kiong Ng. 2016. Learning term embeddings for taxonomic relation identification using dynamic weighting neural network. In *Proceedings of EMNLP*, pages 403–413.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of ACL*, pages 1661–1670.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-embeddings of images and language. In *Proceedings of ICLR (Conference Track)*.

Luke Vilnis and Andrew McCallum. 2015. Word representations via Gaussian embedding. In *Proceedings of ICLR (Conference Track)*.

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.

Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Post-specialisation: Retrofitting vectors of words unseen in lexical resources. In *Proceedings of NAACL-HLT*, pages 516–527.

Ivan Vulić and Nikola Mrkšić. 2018. Specialising word vectors for lexical entailment. In *Proceedings of NAACL-HLT*, pages 1134–1145.

Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. 2014. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING*, pages 2249–2259.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of COLING*, pages 1015–1021.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the ACL*, 3:345–358.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*, pages 1112–1122.

Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. 2014. RC-NET: A general framework for incorporating knowledge into word representations. In *Proceedings of CIKM*, pages 1219–1228.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of ACL*, pages 545–550.

Zheng Yu, Haixun Wang, Xuemin Lin, and Min Wang. 2015. Learning term embeddings for hypernymy identification. In *Proceedings of IJCAI*, pages 1390–1397.

Jingwei Zhang, Jeremy Salwen, Michael Glass, and Alfio Gliozzo. 2014. Word semantic representations using bayesian probabilistic tensor factorization. In *Proceedings of EMNLP*, pages 1522–1531.