Cross-lingual Semantic Specialization via Lexical Relation Induction

Edoardo M. Ponti[1], **Ivan Vulić**[1], **Goran Glavaš**[2], **Roi Reichart**[3], **Anna Korhonen**[1]
[1]Language Technology Lab, TAL, University of Cambridge
[2]Data and Web Science Group, University of Mannheim, Germany
[3]Faculty of Industrial Engineering and Management, Technion, IIT
[1]{ep490,iv250,alk23}@cam.ac.uk
[2]goran@informatik.uni-mannheim.de
[3]roiri@ie.technion.ac.il

Abstract

Semantic specialization integrates structured linguistic knowledge from external resources (such as lexical relations in WordNet) into pretrained distributional vectors in the form of constraints. However, this technique cannot be leveraged in many languages, because their structured external resources are typically incomplete or non-existent. To bridge this gap, we propose a novel method that transfers specialization from a resource-rich source language (English) to virtually *any* target language. Our specialization transfer comprises two crucial steps: 1) Inducing noisy constraints in the target language through automatic word translation; and 2) Filtering the noisy constraints via a state-of-the-art *relation prediction* model trained on the source language constraints. This allows us to specialize any set of distributional vectors in the target language with the refined constraints. We prove the effectiveness of our method through intrinsic word similarity evaluation in 8 languages, and with 3 downstream tasks in 5 languages: lexical simplification, dialog state tracking, and semantic textual similarity. The gains over the previous state-of-art specialization methods are substantial and consistent across languages. Our results also suggest that the transfer method is effective even for lexically distant source-target language pairs. Finally, as a by-product, our method produces lists of WordNet-style lexical relations in resource-poor languages.

1 Introduction

Due to their dependence on the distributional hypothesis (Harris, 1954), that is, word co-occurrence information in large corpora, distributional word embeddings (Mikolov et al., 2013; Levy and Goldberg, 2014; Pennington et al., 2014; Melamud et al., 2016; Bojanowski et al., 2017; Peters et al., 2018, *inter alia*) conflate paradigmatic relations (e.g., synonymy, antonymy, lexical entailment, co-hyponymy, meronymy) and the broader topical (i.e., syntagmatic) relatedness (Schwartz et al., 2015; Mrkšić et al., 2017). This property can propagate undesired effects to language understanding applications such as statistical dialog modeling or text simplification (Faruqui, 2016; Chiu et al., 2016; Mrkšić et al., 2017): for instance, the inability to distinguish between synonymy and antonymy (e.g., between *cheap pubs* and *expensive restaurants*) can break task-oriented dialog or a recommendation system (Mrkšić et al., 2016; Kim et al., 2016b).

Semantic *specialization* techniques are therefore leveraged to stress a relation of interest such as semantic similarity (Wieting et al., 2015; Mrkšić et al., 2017; Ponti et al., 2018) or lexical entailment (Nguyen et al., 2017; Vulić and Mrkšić, 2018) over other types of semantic association in the word vector space. The best-performing specialization models (cf. Mrkšić et al. 2017; Ponti et al. 2018) are executed as vector space post-processors. In short, these techniques force the distributional vectors to conform to external linguistic constraints (e.g., synonymy, meronymy, lexical entailment) extracted from *structured external resources* (e.g., WordNet, BabelNet) to emphasize the particular relation. As post-processors they are applicable to any input distributional space.

A critical requirement for all specialization techniques is the set of linguistic constraints drawn from the curated external semantic resource. Such resources contain incomplete information even in resource-rich languages (e.g., English WordNet), while the resources are scarcer or even non-existent for many other languages. A solution was proposed recently to deal with incomplete information in a resource-rich language: the specialization function learned on the subset of words observed in the external resource gets propagated to the entire vocabulary in a step called *post-specialization* (Vulić

et al., 2018). Yet, another fundamental question concerning specialization techniques is still unresolved: *how to enable specialization in virtually any language, even when the language completely lacks external lexical resources?*

In this work, we therefore propose a novel approach for *cross-lingual specialization transfer* based on Lexical Relation Induction (CLSRI). CLSRI leverages lexical information from a resource-rich language to enable specialization in any target language, without observing a single lexical constraint in the target language. The transfer method consists of two main steps: **1)** We induce a noisy set of constraints in the target language through *automatic word translation* via a shared cross-lingual word vector space (Ruder et al., 2019; Joulin et al., 2018). **2)** To mitigate the noise from the translation process, the initial set of noisy constraints is then refined in a *relation prediction* phase: we adjust a state-of-the-art neural method for lexical relation classification (Glavaš and Vulić, 2018a) and use it to predict the validity of each noisy constraint obtained in the first step. Finally, a standard specialization technique (including the post-specialization step) can then be used *monolingually* in the target language, starting from the set of refined target language constraints.

We verify the usefulness of our specialization transfer method in the intrinsic word similarity task for 8 target languages, followed by 3 downstream tasks in 5 languages: lexical simplification, dialog state tracking, and semantic textual similarity. We observe large improvements over purely distributional word vectors for all target languages and in all tasks. Moreover, we show that the proposed specialization transfer method consistently outperforms the direct specialization transfer based on the composition of the cross-lingual projection and the post-specialization function (Ponti et al., 2018), with substantial gains across all experimental setups. In order to boost the integration of external lexical knowledge into distributional models beyond English, we will release our code and lists of WordNet-style lexical relations generated by our transfer method for all target languages at: `https://github.com/cambridgeltl/xling-postspec`.

2 Related Work

Conflating distinct (both paradigmatic and syntagmatic) lexico-semantic relations is a well-known property of distributional word vectors; semantic specialization of such spaces for a particular lexico-semantic relation (e.g., semantic similarity or lexical entailment) benefits a number of tasks, e.g., dialog state tracking (Mrkšić et al., 2017; Ponti et al., 2018), spoken language understanding (Kim et al., 2016b,a), text simplification (Glavaš and Vulić, 2018b; Ponti et al., 2018), and cross-lingual transfer of resources (Vulić et al., 2017a).

Specialization methods inject external lexical knowledge into a distributional space, tailoring vectors for a particular relation of interest. *Joint specialization* models (Yu and Dredze, 2014; Xu et al., 2014; Bian et al., 2014; Kiela et al., 2015; Liu et al., 2015; Ono et al., 2015; Osborne et al., 2016; Nguyen et al., 2017, *inter alia*) use external constraints to modify the training objective of word embedding models (Mikolov et al., 2013; Dhillon et al., 2015; Liu et al., 2018b,a) and train specialized vectors from scratch.

In contrast, *retrofitting* (also known as *post-processing*) methods tune the pre-trained distributional vectors *post-hoc* based on the provided external constraints. Despite the fact that joint models specialize the entire space, whereas the first generation of retrofitting models specializes only the vectors of words seen in lexical constraints, the latter yield better downstream performance (Mrkšić et al., 2016). Moreover, while the joint models are tightly coupled to a concrete word embedding objective, retrofitting models can be applied on top of any distributional vector space.

Post-specialization (Vulić et al., 2018; Ponti et al., 2018; Kamath et al., 2019) is a generalization of retrofitting that specializes the entire distributional space: 1) it learns a global specialization function using before- and after-retrofitting vectors of words from lexical constraints as training examples and 2) it applies the global specialization functions to vectors of words unseen in lexical constraints. Similar to retrofitting, post-specialization can be applied to any vector space, but also (like *joint* specialization models) specializes the full distributional space.

Since it learns a global and explicit specialization function, post-specialization can be used for cross-lingual specialization transfer. Assuming a shared cross-lingual embedding space (Ruder et al., 2019; Glavaš et al., 2019), a post-specialization function induced on the source language subspace can be directly applied to the target language sub-

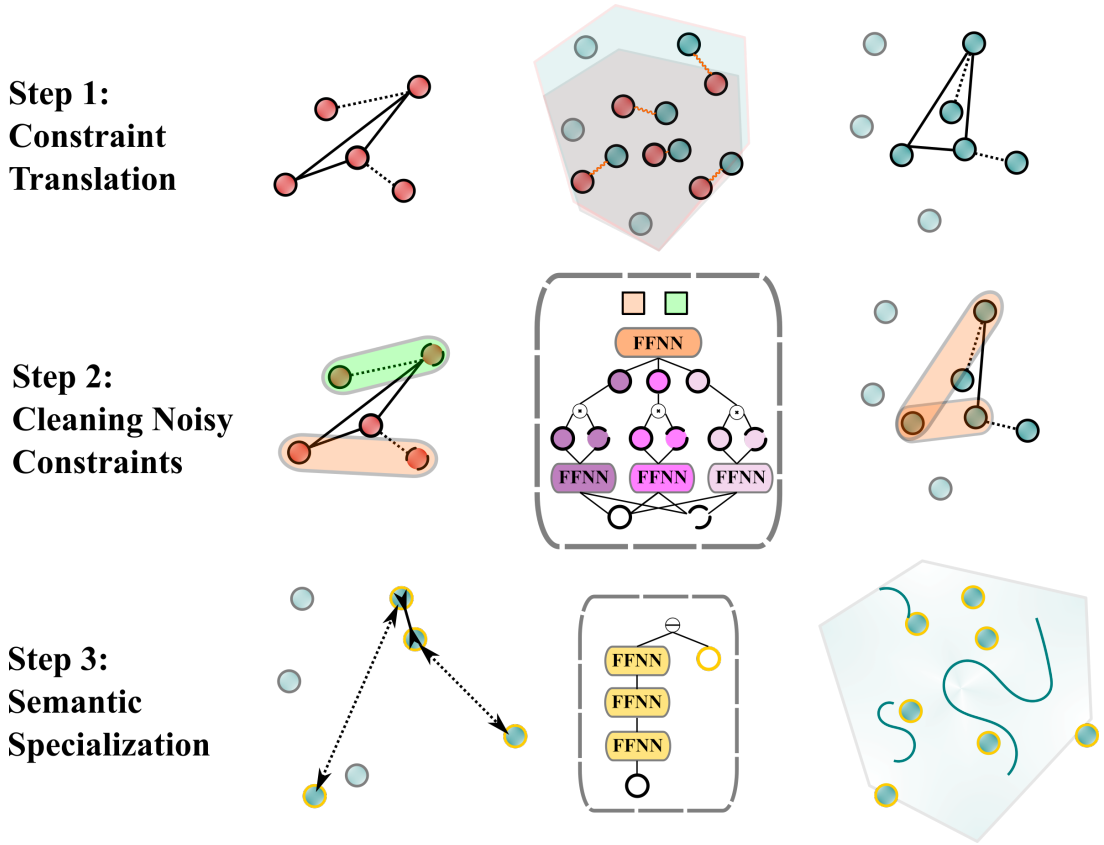

Figure 1: High-level illustration of our CLSRI framework for semantic specialization. Step 1: a network of lexical relations in a source language (red dots, left) is translated into a target language (blue dots, right) through a shared vector space (center). Step 2: a lexical relation classifier (center) trained on vector pairs sampled from the source language (left) prunes the constraints in the noisy target network (right). Step 3: the refined constraints are used to attract or repel the corresponding vectors (golden edges, left); this transformation is learned by a deep feed-forward network (center) and applied to the full target vector space.

space (Glavaš and Vulić, 2018b; Ponti et al., 2018). In this work, we propose a different approach: we use a shared cross-lingual space to (noisily) translate lexical constraints from source to target language, and then use a relation-prediction model (trained on the source language constraints) to filter out the invalid target language constraints. This allows for monolingual application of retrofitting or post-specialization in the target language. Our experiments show that the proposed specialization transfer via lexical relation induction (CLSRI) outperforms the previous state-of-the-art specialization transfer method of Ponti et al. (2018).

3 Methodology

CLSRI in a Nutshell. In cross-lingual semantic specialization our goal is to fine-tune the distributional vectors of a target language L_t leveraging structured knowledge in the form of lexical constraints, available only for a resource-rich source language L_s. To this end, we propose a two-step *translate-and-refine* procedure for the induction of target language constraints, described in § 3.1. We first translate words in each L_s constraint by retrieving their nearest neighbour in L_t from a shared cross-lingual L_s–L_t embedding space (Ruder et al., 2019). Such a translation procedure will generate noisy constraints in the target language due to (1) imperfect word translation via the cross-lingual embedding space and (2) polysemy in L_s and translation of incorrect senses of L_s words. We thus subsequently refine the noisy set of target constraints by having a state-of-the-art neural model for lexico-semantic relation prediction (Glavaš and Vulić, 2018a), trained on the L_s constraints, discern valid from invalid L_t constraints.

Following that, we perform *monolingual retrofitting and post-specialization* in the target language L_t, as outlined in § 3.2. The L_t distributional vectors can be specialized with the cleaned L_t constraints using any off-the-shelf retrofitting model (Faruqui et al., 2015; Mrkšić et al., 2016;

Lengerich et al., 2018, *inter alia*). In this work we opt for the best-performing retrofitting model ATTRACT-REPEL (AR) (Mrkšić et al., 2017; Vulić et al., 2017b). AR specializes only the words seen in the cleaned L_t constraints. As the final step, we generalize AR's specialization to the entire target vocabulary with a *post-specialization* model (Ponti et al., 2018) that learns the global specialization function from pairs of distributional and AR-specialized vectors of words from L_t constraints. A visual summary of our transfer model is presented in Figure 1.

Our proposed CLSRI specialization conceptually differs from an existing cross-lingual specialization transfer methodology (Ponti et al., 2018; Glavaš and Vulić, 2018b), in which the global specialization function is learned in the source language L_s and then transferred directly to the target language L_s via a shared cross-lingual embedding space.

3.1 Induction and Refinement of Constraints

Step 1: Constraint Translation. Following the established methodology of Mrkšić et al. (2017), constraints drawn from external resources are usually split into two broad sets: **1)** ATTRACT constraints couple words that should have similar representations (e.g., synonyms like *complicated* and *complex* or direct hyponym-hypernym pairs like *parrot* and *bird*); and **2)** REPEL constraints indicate which word pairs should appear far-flung in the space (e.g., antonyms like *ancient* and *recent*).

Given a set A_s of ATTRACT word pairs and a set R_s of REPEL word pairs, each word pair (w_s^l, w_s^r) from the vocabulary of the source language $\mathcal{V}_s$ is automatically translated into the target language with vocabulary $\mathcal{V}_t$ using a shared cross-lingual L_s-L_t word embedding space. We create the cross-lingual space $\mathbf{X}_{CL}$ by learning a linear map $\mathbf{W}_{CL}$ that projects the distributional space of the target language $\mathbf{X}_t$ to the distributional space $\mathbf{X}_s$ of the source language, i.e., $\mathbf{X}_{CL} = \mathbf{X}_s \cup \mathbf{X}_t\mathbf{W}_{CL}$. We translate each word w_s from each linguistic constraint in L_s by looking for the nearest neighbour of its vector $\mathbf{x}_s$ in the projected target space $\mathbf{X}_t\mathbf{W}_{CL}$. We employ recently proposed Relaxed Cross-domain Similarity Local Scaling (RCSLS) model of Joulin et al. (2018) to learn the projection matrix $\mathbf{W}_{CL}$ and induce the bilingual space $\mathbf{X}_{CL}$.[1]

Step 2: Cleaning Noisy Constraints. The L_t constraints we obtain by translating L_s constraints via a cross-lingual L_1-L_2 embedding space are expected to be noisy (as validated later in § 5), i.e., a shared cross-lingual space obtained via a linear projection matrix is far from ideal. The translations are going to be particularly noisy for pairs of distant languages for which the projection-based methods for inducing cross-lingual embedding spaces (including RCSLS) generally yield lower bilingual lexicon induction (BLI) performance (Søgaard et al., 2018; Joulin et al., 2018; Glavaš et al., 2019).

In the next step, we therefore clean the noisy L_t constraints obtained via this imperfect translation procedure. To this end, we leverage the state-of-the-art model for lexical relation prediction: the Specialization Tensor Model (STM) (Glavaš and Vulić, 2018a). STM is a neural model that predicts lexical relations for pairs of input distributional vectors based on multi-view projections of those vectors. Each slice of the STM's central specialization tensor specifies a different projection. We modify the original N-ary STM classifier to now model binary classification, and train two instances of the model: one that predicts whether a pair of words represents a valid ATTRACT constraint (A-STM), and another that predicts valid REPEL constraints (R-STM). We train both models with the training instances created from the *clean* L_s constraints.

Given a pair of vectors $(\mathbf{x}_l, \mathbf{x}_r)$ that corresponds to a clean linguistic constraint (w_s^l, w_s^r) from A_s (or R_s), each vector is transformed with k feed-forward networks (FFNs) of the STM model. The paired projections of the two vectors resulting from each FFN are scored with a parameterized biaffine product, producing k latent scores describing the nature of the relation between the input vectors. The k-dimensional latent feature vector is finally passed to a FFN, which performs binary classification.[2] The complete objective is summarized in Equation (1):

$$\text{FFN}^{\sigma} \bigoplus_{i=1}^{k} \left\{ \text{FFN}_i^{\tau}(\mathbf{x}_l)^{\top} \, W_i \, \text{FFN}_i^{\tau}(\mathbf{x}_r) + \mathbf{b}_i \right\} \quad (1)$$

where $\oplus$ stands for concatenation, and the output layer activations are denoted as σ for *sigmoid* and τ for *tanh*.

[1] RCSLS substantially outperforms competing models on the task of bilingual lexicon induction as shown in a recent comparative study (Glavaš et al., 2019), and has been designed to optimize performance exactly on the word translation task.

[2] For further technical details regarding the STM, we refer the reader to the original paper (Glavaš and Vulić, 2018a).

The pairs $(\mathbf{x}_l, \mathbf{x}_r)$ created from A_s and R_s constitute *positive* training instances for A-STM and R-STM, respectively. For each classifier we couple each positive training instance with two types of *negative* training instances: (1) we create a negative instance by substituting a member of the pair ($\mathbf{x}_l$ or $\mathbf{x}_r$) with a randomly sampled vector from one of the other pairs in the same training batch; (2) we create a negative instance by randomly sampling a constraint from the opposing set of constraints, that is, we turn a constraint from A_s into a negative example for R-STM, and, conversely, a constraint from R_s into a negative training instance for A-STM. We train the A-STM and R-STM models with training instances created from L_s constraints and then use the trained model to predict the validity of the translated L_t constraints. We retain only the subsets of L_t constraints A_t and R_t deemed valid by A-STM and R-STM, respectively. Vectors of L_s words (during training) and vectors of L_t words (at inference) are taken from the induced bilingual L_s-L_t space $\mathbf{X}_{CL} = \mathbf{X}_s \cup \mathbf{X}_t\mathbf{W}_{CL}$.

3.2 Semantic Specialization

We can now directly feed A_t and R_t to any retrofitting model and (monolingually) specialize any distributional space in the target language. We first run the state-of-the-art retrofitting model ATTRACT-REPEL (AR) (Mrkšić et al., 2017) with A_t and R_t constraints. AR however, specializes only the words present in A_t and R_t. In the next step, we generalize AR's specialization to the full vocabulary $\mathcal{V}_t$ with the state-of-the-art post-specialization model (Ponti et al., 2019). For completeness, we briefly summarize AR and the post-specialization model of Ponti et al. (2019).

Retrofitting with ATTRACT-REPEL. Each constraint from A_t and R_t is used to fine-tune the distance between their corresponding vectors $(\mathbf{x}_l, \mathbf{x}_r)$ in the target L_t distributional space. Let $\mathcal{B}_A$ be a batch of vector pairs created from ATTRACT constraints A_t and $\mathcal{B}_R$ the batch of vector pairs created from REPEL constraints R_t. For each batch $\mathcal{B}_A$ and each batch $\mathcal{B}_R$, we construct batches of corresponding negative pairs $\mathcal{T}_A(\mathcal{B}_A)$ and $\mathcal{T}_R(\mathcal{B}_R)$, containing new pairs of words sampled among those present in the batch of positive pairs. In particular, half of the negative examples $\mathbf{t}_l$ and $\mathbf{t}_r$ for ATTRACT (or REPEL) pairs are chosen by retrieving the nearest (or farthest) neighbours to $\mathbf{x}_l$ and $\mathbf{x}_r$, respectively, in terms of cosine similarity. Another half are random negative examples.

AR minimizes an objective based on max-margin loss between positive pairs and their corresponding negative pairs. More precisely, its objective has three loss components: $\mathcal{L}_{AR} = Att(\mathcal{B}_A, \mathcal{T}_A) + Rep(\mathcal{B}_R, \mathcal{T}_R) + Pre(\mathcal{B}_A, \mathcal{B}_R)$. The first component ensures that word pairs from each $\mathcal{B}_A$ are drawn closer together than those in the corresponding T_A up to a certain "attract" margin δ_A:

$$Att(\mathcal{B}_A, \mathcal{T}_A) = \sum_{i=1}^{|\mathcal{B}_A|} \Big[\tau\big(\delta_A \mathbf{x}_l^i \mathbf{t}_l^i - \mathbf{x}_l^i \mathbf{x}_r^i\big) + \tau\big(\delta_A + \mathbf{x}_r^i \mathbf{t}_r^i - \mathbf{x}_l^i \mathbf{x}_r^i\big)\Big] \quad (2)$$

where $\tau(z) = \max(0, z)$ is ramp function. Analogously, $Rep(\mathcal{B}_R, \mathcal{T}_R)$ forces the vectors of words in $\mathcal{B}_R$ pairs to be further away than the vectors of their corresponding $\mathcal{T}_R$ pairs by a margin δ_R. Finally, $Pre(\mathcal{B}_A, \mathcal{B}_R)$ is the regularization objective that preserves the useful semantic information from the distributional space by minimizing the Euclidean distance between original and changed vectors.[3]

Post-Specialization. By virtue of AR retrofitting, only the subset of vectors of L_t words *observed* in the refined L_t constraints are specialized. The specialized subspace, however, contains useful information for propagating the specialization to the rest of the vocabulary $\mathcal{V}_t$ (i.e., to the vectors of L_t words *unseen* in A_t and R_t). Post-specialization aims to learn a global specialization function $G : \mathbf{X}_t \in \mathbb{R}^d \rightarrow \mathbf{X}'_t \in \mathbb{R}^d$ that approximates the perturbation patterns of AR as captured by changes in vectors of *seen* words from A_t and R_t. G is learned as a non-linear mapping between pairs ($\mathbf{x}_i$, $\mathbf{y}_i$), where $\mathbf{x}_i \in \mathbf{X}_t$ is a distributional vectors of some constraint word (from A_t or R_t) and $\mathbf{y}_i$ is its corresponding AR-specialized vector. In line with Vulić et al. (2018) and Ponti et al. (2018), we implement this function as a deep feed-forward neural network with *l* hidden layers of size *h* and a final linear layer with weight $W \in \mathcal{R}^{h \times d}$. We optimize model parameters θ_G by minimizing a contrastive margin ranking loss with random confounders (Weston et al., 2011, *inter alia*). The cosine similarity between a distributional vector transformed with G and the corresponding "gold" vector (i.e., AR-specialized vector) is forced to be larger than that

[3] For more details about the AR objectives, we refer the interested reader to the original work (Mrkšić et al., 2017).

between the former and randomly sampled confounders (k of them) by a margin δ_{MM}:

$$\sum_{i=1}^{||\mathcal{V}_s||} \sum_{j=1|j\neq i}^{k} \tau[\delta_{MM} - \cos(G(\mathbf{x}_i;\theta_G),\mathbf{y}_i) + \\ + \cos(G(\mathbf{x}_i;\theta_G),\mathbf{y}_j)] = \mathcal{L}_{MM} \quad (3)$$

Once the global specialization transformation G is learned, it is applied to the whole distributional space of our target language: $\mathbf{Y}_t = G_{\theta_G}(\mathbf{X}_t)$.

Note that with our proposed specialization approach CLSRI, we execute the retrofitting and post-specialization completely monolingually in the target language L_t on the automatically induced constraints in the target language. In contrast, existing work (Vulić et al., 2018; Glavaš and Vulić, 2018b; Ponti et al., 2018) transfers the post-specialization function learned for the source language L_s to the target language L_t via a cross-lingual vector space. This fundamental design difference is illustrated in Figure 1 and empirically validated in §5.

4 Experimental Setup

Lexical Constraints. The assortment of English constraints for specialization is the same as in prior work (Zhang et al., 2014; Ono et al., 2015; Vulić et al., 2018; Ponti et al., 2018). These constraints concern the lexical relations documented in WordNet (Fellbaum, 1998) and Roget's Thesaurus (Kipfer, 2009). Initially, they amount to 1,023,082 synonymy/ATTRACT word pairs and 380,873 antonymy/REPEL pairs, which cover 14.6% of the 200K most frequent English words, as found in the vocabulary of FASTTEXT vectors (Bojanowski et al., 2017). The number of constraints is substantially reduced in the target languages[4] after the induction process from § 3.1, both after the rough translation and after the refinement via relation prediction. The actual numbers are reported in Figure 2.

Distributional Word Vectors. In order to scale up our evaluation to a representative sample of languages and language types (O'Horan et al., 2016; Ponti et al., 2019), we use 300-dim distributional vectors from the FASTTEXT[5] collection

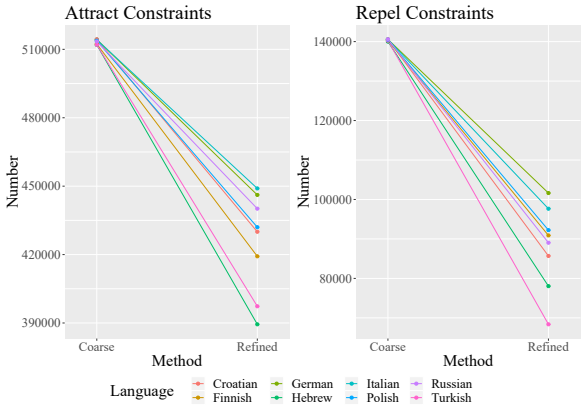

Figure 2: Number of constraints in the target languages after the initial coarse translation (*Coarse*) and after relation prediction (*Refined*).

(Bojanowski et al., 2017) trained on Wikipedia, which offers off-the-shelf comparable models for hundreds of languages. The vectors of each language are aligned to English using RCSLS (Joulin et al., 2018) (see § 3.1),[6] with the suggested hyper-parameters: 10 nearest neighbours in the RCSLS loss function, 10 iterations, and a language pair-dependent learning rate tuned from the set $\{1, 10, 25, 50\}$.

Relation Prediction. The STM model is trained with the Adam optimizer (Kingma and Ba, 2015), a learning rate of 0.0001, and a batch size of 48 (including negative examples) for a maximum of 10 iterations. Early stopping was implemented based on the F_1 score on a development set comprising 5% of the source language constraints. The hidden layer dimensionality is 300, and we use $k = 5$ specialization sub-tensors. Regarding the quality of the STM predictions, the best models achieve an F_1 score of 81.4 on ATTRACT constraints, and an F_1 score of 66.9 on REPEL constraints.[7]

AR and Post-Specialization. We retain the exact hyper-parameter configuration for ATTRACT-REPEL from the original work (Mrkšić et al., 2017): $\delta_A = 0.6$, $\delta_R = 0.0$, $\lambda_P = 10^{-9}$. Adagrad (Duchi et al., 2011) is employed to optimize the model parameters for 5 epochs, feeding batches of size $|\mathcal{B}_A| = |\mathcal{B}_R| = 50$, again as in prior work.

[4] These include Croatian (HR), Finnish (FI), German (DE), Hebrew (HE), Italian (IT), Polish (PL), Russian (RU), and Turkish (TR). We also use Portuguese (PT), Spanish (ES), and Arabic (AR) in some evaluations. The languages were selected according to the data availability in our evaluation tasks.

[5] FASTTEXT is trained on Wikipedia through an adaptation of Skip-Gram with Negative Sampling (SGNS) that builds representations for each word's constituent character n-grams and sums them up to obtain the entire word's representation.

[6] https://github.com/facebookresearch/fastText/tree/master/alignment

[7] Although the predicted REPEL constraints are quite inaccurate, we verified empirically that excluding them would harm the overall performance.

Owing to the difference in the amount of supervision, the post-specialization model has partially non-overlapping configurations for the baseline model of Ponti et al. (2018) and our CLSRI model. For both models each of the $l = 3$ hidden layers of the feed-forward network is composed of $h = 2,048$ hidden units, and is non-linearly activated by LeakyReLU (Maas et al., 2013). We apply a dropout of 0.2 both in input and between hidden layers. In Eq. (3), the margin $\delta_{MM} = 1$, and the negative examples amount to $k = 25$. We use SGD with the learning rate $lr = 0.1$. For the baseline, post-specialization is trained for 10 epochs, each consisting of 1 million mini-batches of 32 randomly sampled pairs. For our CLSRI model, it is limited to 2 epochs of 200K iterations each.[8]

Models in Comparison. Finally, we summarize the main models benchmarked in §5. First, we evaluate the original **Distributional** vectors. **X-PS** refers to the baseline model of Ponti et al. (2018) based on direct cross-lingual post-specialization. **CLSRI-AR** denotes the variant of our model based on constraint induction in L_t after running the initial AR retrofitting without post-specialization; **CLSRI-PS** refers to our full model with the post-specialization step.

5 Results and Discussion

We evaluate different specialization models across several target languages on the intrinsic word similarity task and three downstream language understanding tasks where distinguishing between true semantic similarity and conceptual relatedness is crucial: dialog state tracking, lexical simplification, and semantic textual similarity. The choice of tasks has also been driven by the availability of standardized evaluation data in different languages.

5.1 Word Similarity

Evaluation Setup. The intrinsic evaluation is based on a set of (true) word similarity benchmarks manually translated from (subsets of) the English SimLex (Hill et al., 2015) and re-scored in the target languages.[9] In particular, the benchmarks are collected from the work of Leviant and Reichart (2015) for German, Italian, and Russian (999 pairs),[10] from Mrkšić et al. (2017) for Hebrew and Croatian (999 pairs),[11] from Venekoski and Vankka (2017) for Finnish (300),[12] from Mykowiecka et al. (2018) for Polish (999),[13] and from Ercan and Yıldız (2018) for Turkish (500).[14] We measure the Spearman's ρ rank correlation between the gold human-elicited word pair similarity scores and the cosine similarity of the corresponding word vectors retrieved from each vector space.

Results and Analysis. We summarize the results for word similarity in Table 1. The full CLSRI-PS model outperforms both the distributional vectors and the baseline method for cross-lingual specialization (Ponti et al., 2018). In all languages but two (DE and RU) even the CLSRI-AR model without post-specialization is superior to both baselines, and the post-specialization step additionally improves the results, supporting the findings from prior work (Vulić et al., 2018). Crucially, the performance of CLSRI-PS remains strong even for distant language pairs (e.g., for EN–HE, EN–TR or EN–FI), whereas the X-PS baseline shows a drop in performance for such cases. We suspect that it is because the success of our CLSRI-PS method depends less on the quality of the underlying shared cross-lingual vector space, which is known to deteriorate for more distant language pairs (Søgaard et al., 2018; Glavaš et al., 2019).

5.2 Dialog State Tracking

A standard language understanding evaluation task used in prior work on semantic specialization (Mrkšić et al., 2017; Ponti et al., 2018, *inter alia*) is dialog state tracking (DST) (Henderson et al., 2014; Mrkšić et al., 2017). A DST model is a fundamental building block of statistical modular dialogue systems (Young, 2010). Its task is to maintain the information of the user's goals during a multi-turn conversation by updating the dialog belief state at each turn. Distinguishing true similarity as captured in specialized word vectors from broader relatedness is crucial for DST to succeed: e.g., a dialog system for restaurant bookings should not confuse the *western* and the *eastern* part of town, or *Thai* and *Japanese* cuisine.

[8] For both models, the hyper-parameters are chosen with a grid search over the intervals h={1024, 2048, 4096}, l={2, 3}, lr={0.1, 0.01, 0.001}, and optimizers in {*Adam, SGD*}, using a held-out dev set (10% of the constraints).

[9] In contrast to other datasets like WordSim-353 (Finkelstein et al., 2002) or MEN (Bruni et al., 2014), SimLex encourages scores to distinguish between pure semantic similarity (actual synonyms) and broad topical relatedness.

[10] http://leviants.com/ira.leviant/MultilingualVSMdata.html

[11] https://github.com/nmrksic/attract-repel

[12] https://github.com/venekoski/FinSemEvl

[13] http://zil.ipipan.waw.pl/CoDeS

[14] http://www.gokhanercan.com/resources/anlamver.aspx

Model	DE	IT	HE	FI	HR	TR	PL	RU
Distributional	.426	.304	.368	.240	.344	.535	.395	.270
X-PS	.503	.392	.380	.314	.376	.464	.344	.402
CLSRI-AR	.500	**.525**	.454	.394	.425	.554	.433	.331
CLSRI-PS	**.565**	.512	**.522**	**.490**	**.505**	**.613**	**.534**	**.507**

Table 1: Spearman's ρ correlation scores for 8 languages on datasets for intrinsic evaluation of true semantic similarity. The models in comparison are briefly summarized in § 4 and in Figure 1.

Model	DE	IT
Distributional	0.640	0.729
X-Postspec	0.647	0.737
CI-AR	0.652	0.745
CI-Postspec	**0.687**	**0.782**

Table 2: Joint goal accuracy scores in the DST task.

Evaluation Setup. To be directly comparable to prior work when evaluating the effects of specialized word embeddings on DST, we rely on the Neural Belief Tracker (NBT) v2 (Mrkšić and Vulić, 2018): it is a fully statistical DST model that operates solely on the basis of pretrained word vectors (Mrkšić et al., 2017), and they are pivotal to its performance.[15] Again following prior work, our evaluation data come from the multilingual Wizard-of-Oz (WOZ) dataset (Wen et al., 2017), which is available in two target languages: German and Italian (Mrkšić et al., 2017). It contains 1,200 dialogues split into training (600 dialogues), development (200), and test data (400). We report the standard DST metric of *joint goal accuracy*: it refers to the proportion of dialog turns where all the users goals were correctly identified.

Results and Analysis. The results on the German and Italian DST task are summarized in Table 2. Several findings emerge from the results. First, as already confirmed in prior work (Vulić et al., 2018; Ponti et al., 2018), vectors specialized for semantic similarity are indeed important for DST: we observe improvements with all specialized vectors. The highest gains are observed with the full CSLRI-PS model. This confirms two main intuitions: 1) our proposed specialization transfer via lexical induction in the target language is more robust than the previous X-PS method of Ponti et al. (2018), and 2) the full-vocabulary post-specialization step is again useful as the initial CSLRI-AR model cannot match the performance of CSLRI-PS.

[15]Note that the original NBT framework in the English DST task has been recently surpassed by more intricate task-specific architectures (Zhong et al., 2018; Ren et al., 2018), but its lightweight design coupled with its strong dependence on input word vectors still makes it a convenient means to evaluate the effects of different specialization methods.

5.3 Lexical Simplification

Lexical simplification (LS) aims to automatically replace complex words (i.e., specialized terms, words used less frequently and known to fewer speakers) with their simpler in-context synonyms: the simplified text must be grammatical and retain the meaning of the original text. Lexical simplification critically depends on discerning semantic similarity from other types of semantic relatedness, as the meaning of the original text might not be preserved otherwise (e.g., *"The orange automobile crashed."* vs. *"The orange wheel crashed."*).

Evaluation Setup. To evaluate the effects of similarity-based specialization on LS, we employ Light-LS (Glavaš and Štajner, 2015), a language-agnostic LS tool that makes simplifications based on word similarities in a given vector space. The quality of similarity-based information encoded in the vector space encode is thus expected to directly correlate with the performance of Light-LS. We use LS datasets for Italian (IT) (Tonelli et al., 2016), Spanish (ES) (Saggion et al., 2015; Saggion, 2017), and Portuguese (PT) (Hartmann et al., 2018) to evaluate the specialized spaces in those languages. We rely on the standard LS evaluation metric of *Accuracy* (Horn et al., 2014; Glavaš and Štajner, 2015): it quantifies both the quality and frequency of replacements as a number of correct simplifications divided by the total number of complex words.

Results and Analysis. The results are reported in Table 3. As shown in previous work (Vulić et al., 2018; Ponti et al., 2018), retrofitting (CLSRI-AR) and the cross-lingual post-specialization transfer (X-PS) are substantially better in the LS task than the original distributional space. However, our full CLSRI-PS model results in substantial boosts in the

Model	LS			STS
	IT	ES	PT	AR
Distributional	0.28	0.27	0.27	0.67
X-PS	0.38	0.57	0.55	0.66
CLSRI-AR	0.35	0.51	0.33	0.66
CLSRI-PS	**0.51**	**0.74**	**0.72**	**0.70**

Table 3: Lexical Simplification (LS) performance for Italian, Spanish, and Portuguese; Semantic Textual Similarity (STS) performance for Arabic.

LS task (13-17%) over the previous best reported scores of X-PS as well as over CLSRI-AR.

5.4 Semantic Text Similarity

Evaluation Setup. Finally, we also carry downstream evaluation in the semantic textual similarity (STS) task. The Arabic dataset constructed for SemEval 2017 track 1[16] (Cer et al., 2017) consists of sentence pairs scored from 0 (semantic independence) to 5 (semantic equivalence). We augment the training set with all the data for English (translated with Google Translate) from previous editions of the shared task. To classify sentence pairs, we employ the CNN-HTCI model (Shao, 2017). Each sentence is encoded with a convolutional network into a hidden representation. Then, the interaction between the pair of representations is evaluated as their element-wise multiplication and absolute difference. A fully connected network takes this interaction as input, and infers the similarity score.

Results and Analysis. We report the accuracy scores for the Arabic STS in Table 3. Interestingly, for STS both X-PS and CLSRI-AR damage the performance of the distributional baseline. However, the full CLSRI-PS model still shows a substantial improvement over all baselines. This again suggests its wide stability and effectiveness.

To empirically validate the importance of noisy constraints refinement (see § 3.1), we have also evaluated an ablated variant of CLSRI-PS without the refinement step: this model variant relies only on noisy translations of L_s lexical constraints. While this variant leads to improvements over the X-PS baseline across the board, it is consistently outperformed by the full CSLRI-PS model in downstream tasks: e.g., the gains with the full model are 2-3% in the LS task, and 2% in the Arabic STS task. Since the full CSLRI-PS model does not require any additional input for the lexical prediction step (i.e., it operates with the same set of L_s constraints as the translation step), these results suggest that both steps should be applied for improved specialization in the target languages.

[16] http://alt.qcri.org/semeval2017/task1/

6 Future Work

As a supplemental benefit of CLSRI, the constraints induced by translation and pruning hold promise to create WordNet-style resources for languages that lack structured linguistic knowledge. While the relations extracted in this proof-of-concept paper do not cover the rich and expressive set of WordNet relations in its entirety, they are nonetheless sufficient to create parts of the core WordNet structure with synsets (synonyms) and lexical relations across synsets (antonyms) from scratch. Notably, however, our method operates at the word level, rather than the sense level.

Furthermore, our method is amenable be extended to other WordNet lexical relations sych as hypernyms and hyponyms. In recent works, procedures of retrofitting (Vulić and Mrkšić, 2018) and post-specialization (Kamath et al., 2019) have been developed for lexical entailment. These procedures can be easily adapted to the semantic specialization step presented in § 3.2, whereas constraint translation and refinement (§ 3.1) are relation-agnostic. We will exploit this direction in future work.

7 Conclusion

We have proposed a new method for cross-lingual transfer of semantic specialization via induction of lexical constraints in a resource-poor target language. We have verified its usefulness in intrinsic and extrinsic language understanding tasks and across a spectrum of target languages. We report consistent improvements over previous state-of-the-art specialization methods. Crucially, our method is robust to target languages that are distant from source languages, as its performance is consistent across all considered language pairs.

Acknowledgements

This work is supported by the ERC Consolidator Grant LEXICAL (no 648909). The work of Goran Glavaš is supported by the Baden-Württemberg Stiftung (AGREE grant of the Eliteprogramm). We thank the three anonymous reviewers for their helpful comments and suggestions.

References

Jiang Bian, Bin Gao, and Tie-Yan Liu. 2014. Knowledge-powered deep learning for word embedding. In *Proceedings of ECML-PKDD*, pages 132–148.

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

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

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of SEMEVAL*, pages 1–14.

Billy Chiu, Anna Korhonen, and Sampo Pyysalo. 2016. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of REPEVAL*, pages 1–6.

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

John C. Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Gökhan Ercan and Olcay Taner Yıldız. 2018. Anlamver: Semantic model evaluation dataset for turkish-word similarity and relatedness. In *Proceedings of COLING*, pages 3819–3836.

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.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Goran Glavaš and Ivan Vulić. 2018a. Discriminating between lexico-semantic relations with the specialization tensor model. In *Proceedings of NAACL-HLT*, pages 181–187.

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

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of ACL*, pages 710–721.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of ACL*, pages 63–68.

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

Nathan S. Hartmann, Gustavo H. Paetzold, and Sandra M. Aluísio. 2018. SIMPLEX-PB: A lexical simplification database and benchmark for Portuguese. In *Proceedings of the 2018 International Conference on Computational Processing of the Portuguese Language*, pages 272–283.

Matthew Henderson, Blaise Thomson, and Steve Young. 2014. Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation. In *Proceedings of IEEE SLT*, pages 360–365.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a lexical simplifier using Wikipedia. In *Proceedings of ACL*, pages 458–463.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of EMNLP*, pages 2979–2984.

Aishwarya Kamath, Jonas Pfeiffer, Edoardo Maria Ponti, Goran Glavaš, and Ivan Vulić. 2019. Specializing distributional vectors of all words for lexical entailment. In *Proceedings of the 4th Workshop on Representation Learning for NLP*, pages 72–83.

Douwe Kiela, Felix Hill, and Stephen Clark. 2015. Specializing word embeddings for similarity or relatedness. In *Proceedings of EMNLP*, pages 2044–2048.

Joo-Kyung Kim, Marie-Catherine de Marneffe, and Eric Fosler-Lussier. 2016a. Adjusting word embeddings with semantic intensity orders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 62–69.

Joo-Kyung Kim, Gokhan Tur, Asli Celikyilmaz, Bin Cao, and Ye-Yi Wang. 2016b. Intent detection using semantically enriched word embeddings. In *Proceedings of SLT*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR (Conference Track)*.

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

Benjamin J. Lengerich, Andrew L. Maas, and Christopher Potts. 2018. Retrofitting distributional embeddings to knowledge graphs with functional relations. In *Proceedings of COLING*, pages 2423–2436.

Ira Leviant and Roi Reichart. 2015. Separated by an un-common language: Towards judgment language informed vector space modeling. *arXiv preprint arXiv:1508.00106*.

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

Qian Liu, Heyan Huang, Yang Gao, Xiaochi Wei, Yuxin Tian, and Luyang Liu. 2018a. Task-oriented word embedding for text classification. In *Proceedings of COLING*, pages 2023–2032.

Qian Liu, Heyan Huang, Guangquan Zhang, Yang Gao, Junyu Xuan, and Jie Lu. 2018b. Semantic structure-based word embedding by incorporating concept convergence and word divergence. In *Proceedings of AAAI*, pages 5261–5268.

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. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of ICML*.

Oren Melamud, David McClosky, Siddharth Patwardhan, and Mohit Bansal. 2016. The role of context types and dimensionality in learning word embeddings. In *Proceedings of NAACL-HLT*, pages 1030–1040.

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 NeurIPS*, pages 3111–3119.

Nikola Mrkšić and Ivan Vulić. 2018. Fully statistical neural belief tracking. In *Proceedings of ACL*, pages 108–113.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of ACL*, pages 1777–1788.

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.

Agnieszka Mykowiecka, Malgorzata Marciniak, and Piotr Rychlik. 2018. SimLex-999 for Polish. In *Proceedings of LREC*, pages 2398–2402.

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.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING*, pages 1297–1308.

Masataka Ono, Makoto Miwa, and Yutaka Sasaki. 2015. Word embedding-based antonym detection using thesauri and distributional information. In *Proceedings of NAACL-HLT*, pages 984–989.

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

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, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*.

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.

Liliang Ren, Kaige Xie, Lu Chen, and Kai Yu. 2018. Towards universal dialogue state tracking. In *Proceedings of EMNLP*, pages 2780–2786.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Horacio Saggion. 2017. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for Spanish. *ACM Transactions on Accessible Computing*, 6(4):14.

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.

Yang Shao. 2017. HCTI at SemEval-2017 Task 1: Use convolutional neural network to evaluate semantic textual similarity. In *Proceedings of SEMEVAL*, pages 130–133.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of ACL*, pages 778–788.

Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. SIMPITIKI: A simplification corpus for Italian. In *Proceedings of CLiC-it*.

Viljami Venekoski and Jouko Vankka. 2017. Finnish resources for evaluating language model semantics. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 231–236.

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.

Ivan Vulić, Nikola Mrkšić, and Anna Korhonen. 2017a. Cross-lingual induction and transfer of verb classes based on word vector space specialisation. In *Proceedings of EMNLP*, pages 2536–2548.

Ivan Vulić, Nikola Mrkšić, Roi Reichart, Diarmuid Ó Séaghdha, Steve Young, and Anna Korhonen. 2017b. Morph-fitting: Fine-tuning word vector spaces with simple language-specific rules. In *Proceedings of ACL*, pages 56–68.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of EACL*.

Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. WSABIE: Scaling up to large vocabulary image annotation. In *Proceedings of IJCAI*, pages 2764–2770.

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.

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.

Steve Young. 2010. Still talking to machines (cognitively speaking). In *Proceedings of INTERSPEECH*, pages 1–10.

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

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.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of ACL*, pages 1458–1467.