Semantic Specialization of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints

Nikola Mrkšić[1,2], Ivan Vulić[1], Diarmuid Ó Séaghdha[2], Ira Leviant[3]
Roi Reichart[3], Milica Gašić[1], Anna Korhonen[1], Steve Young[1,2]
[1] University of Cambridge
[2] Apple Inc.
[3] Technion, IIT

Abstract

We present ATTRACT-REPEL, an algorithm for improving the semantic quality of word vectors by injecting constraints extracted from lexical resources. ATTRACT-REPEL facilitates the use of constraints from mono- and cross-lingual resources, yielding semantically specialized cross-lingual vector spaces. Our evaluation shows that the method can make use of existing cross-lingual lexicons to construct high-quality vector spaces for a plethora of different languages, facilitating semantic transfer from high- to lower-resource ones. The effectiveness of our approach is demonstrated with state-of-the-art results on semantic similarity datasets in six languages. We next show that ATTRACT-REPEL-specialized vectors boost performance in the downstream task of dialogue state tracking (DST) across multiple languages. Finally, we show that cross-lingual vector spaces produced by our algorithm facilitate the training of multilingual DST models, which brings further performance improvements.

1 Introduction

Word representation learning has become a research area of central importance in modern natural language processing. The common techniques for inducing distributed word representations are grounded in the distributional hypothesis, relying on co-occurrence information in large textual corpora to learn meaningful word representations (Mikolov et al., 2013b; Pennington et al., 2014; Ó Séaghdha and Korhonen, 2014; Levy and Goldberg, 2014). Recently, methods that go beyond stand-alone unsupervised learning have gained increased popularity. These models typically build on distributional ones by using human- or automatically-constructed knowledge bases to enrich the semantic content of existing word vector collections. Often this is done as a post-processing step, where the distributional word vectors are refined to satisfy constraints extracted from a lexical resource such as WordNet (Faruqui et al., 2015; Wieting et al., 2015; Mrkšić et al., 2016). We term this approach *semantic specialization*.

In this paper we advance the semantic specialization paradigm in a number of ways. We introduce a new algorithm, ATTRACT-REPEL, that uses synonymy and antonymy constraints drawn from lexical resources to tune word vector spaces using linguistic information that is difficult to capture with conventional distributional training. Our evaluation shows that ATTRACT-REPEL outperforms previous methods which make use of similar lexical resources, achieving state-of-the-art results on two word similarity datasets: SimLex-999 (Hill et al., 2015) and SimVerb-3500 (Gerz et al., 2016).

We then deploy the ATTRACT-REPEL algorithm in a multilingual setting, using semantic relations extracted from BabelNet (Navigli and Ponzetto, 2012; Ehrmann et al., 2014), a cross-lingual lexical resource, to inject constraints between words of different languages into the word representations. This allows us to embed vector spaces of multiple languages into a single vector space, exploiting information from high-resource languages to improve the word representations of lower-resource ones. Table 1 illustrates the effects of cross-lingual ATTRACT-REPEL specialization by showing the nearest neighbors for three English words across three cross-lingual spaces.

Transactions of the Association for Computational Linguistics, vol. 5, pp. 309–324, 2017. Action Editor: Sebastian Padó.
Submission batch: 4/2017; Revision batch: 7/2017; Published 9/2017.

en_morning			en_carpet			en_woman		
Slavic+EN	**Germanic**	**Romance+EN**	**Slavic+EN**	**Germanic**	**Romance+EN**	**Slavic + EN**	**Germanic**	**Romance+EN**
en_daybreak	de_vormittag	pt_madrugada	en_rug	de_teppichboden	en_rug	ru_женщина	de_frauen	fr_femme
en_morn	nl_krieken	it_mattina	bg_килим	nl_tapijten	it_moquette	bg_жените	sv_kvinnliga	en_womanish
bg_разсъмване	en_dawn	en_dawn	ru_ковролин	en_rug	it_tappeti	hr_žena	sv_kvinna	es_mujer
hr_svitanje	nl_zonsopkomst	pt_madrugadas	bg_килими	de_teppich	pt_tapete	en_womanish	sv_kvinnor	pt_mulher
hr_zore	sv_morgonen	es_madrugada	pl_dywany	en_carpeting	es_moqueta	bg_жена	de_weib	es_fémina
bg_изгрев	de_tagesanbruch	it_nascente	bg_мокет	de_teppiche	it_tappetino	pl_kobieta	en_womanish	en_womens
en_dawn	en_sunrise	en_morn	pl_dywanów	sv_mattor	en_carpeting	hr_treba	sv_kvinno	pt_feminina
ru_утро	nl_opgang	es_aurora	hr_tepih	sv_matta	pt_carpete	bg_жени	de_frauenzimmer	pt_femininas
bg_аврора	de_sonnenaufgang	fr_matin	pl_wykładziny	en_carpets	pt_tapetes	en_womens	sv_honkön	es_femina
hr_jutro	nl_dageraad	fr_aurora	ru_ковер	nl_tapijt	fr_moquette	pl_kobiet	sv_kvinnan	fr_femelle
ru_рассвет	de_anbruch	es_amaneceres	ru_коврик	nl_kleedje	en_carpets	hr_žene	nl_vrouw	pt_fêmea
hr_zora	sv_morgon	en_sunrises	hr_ćilim	nl_vloerbedekking	es_alfombra	pl_niewiasta	de_madam	fr_femmes
hr_zoru	en_daybreak	es_mañanero	en_carpeting	de_brücke	es_alfombras	hr_žensko	sv_kvinnligt	it_donne
pl_poranek	de_morgengrauen	fr_matinée	pl_dywan	de_matta	fr_tapis	hr_ženke	sv_gumman	es_mujeres
en_sunrise	nl_zonsopgang	it_mattinata	ru_ковров	nl_matta	pt_tapeçaria	pl_samica	sv_female	pt_fêmeas
bg_зазоряване	nl_goedemorgen	pt_amanhecer	en_carpets	en_mat	it_zerbino	ru_самка	sv_gumma	es_hembras
bg_сутрин	sv_gryningen	en_cockcrow	ru_килим	de_matte	it_tappeto	bg_женска	sv_kvinnlig	en_wife
en_sunrises	en_mornin	pt_aurora	en_mat	en_doilies	es_tapete	hr_ženka	sv_feminin	fr_nana
bg_зора	sv_gryning	pt_alvorecer	hr_sag	nl_mat	es_manta	ru_дама	en_wife	es_hembra

Table 1: Nearest neighbors for three example words across Slavic, Germanic and Romance language groups (with English included as part of each word vector collection). Semantically dissimilar words have been underlined.

In each case, the vast majority of each words' neighbors are meaningful synonyms/translations.[1]

While there is a considerable amount of prior research on joint learning of cross-lingual vector spaces (see Section 2.2), to the best of our knowledge we are the first to apply semantic specialization to this problem.[2] We demonstrate its efficacy with state-of-the-art results on the four languages in the Multilingual SimLex-999 dataset (Leviant and Reichart, 2015). To show that our approach yields semantically informative vectors for lower-resource languages, we collect intrinsic evaluation datasets for Hebrew and Croatian and show that cross-lingual specialization significantly improves word vector quality in these two (comparatively) low-resource languages.

In the second part of the paper, we explore the use of ATTRACT-REPEL-specialized vectors in a downstream application. One important motivation for training word vectors is to improve the lexical coverage of supervised models for language understanding tasks, e.g., question answering (Iyyer et al., 2014) or textual entailment (Rocktäschel et al., 2016). In this work, we use the task of *dialogue state tracking* (DST) for extrinsic evaluation. This task, which arises in the construction of statistical dialogue systems (Young et al., 2013), involves understanding the goals expressed by the user and updating the system's distribution over such goals as the conversation progresses and new information becomes available.

We show that incorporating our specialized vectors into a state-of-the-art neural-network model for DST improves performance on English dialogues. In the multilingual spirit of this paper, we produce new Italian and German DST datasets and show that using ATTRACT-REPEL-specialized vectors leads to even stronger gains in these two languages. Finally, we show that our cross-lingual vectors can be used to train a single model that performs DST in all three languages, in each case outperforming the monolingual model. To the best of our knowledge, this is the first work on multilingual training of any component of a statistical dialogue system. Our results indicate that multilingual training holds great promise for bootstrapping language understanding models for other languages, especially for dialogue domains where data collection is very resource-intensive.

All resources related to this paper are available at `www.github.com/nmrksic/attract-repel`. These include: **1)** the ATTRACT-REPEL source code; **2)** bilingual word vector collections combining English with 51 other languages; **3)** Hebrew and Croatian intrinsic evaluation datasets; and **4)** Italian and German Dialogue State Tracking datasets collected for this work.

[1] Some (negative) effects of the distributional hypothesis do persist. For example, *nl_krieken* (Dutch for *cherries*), is identified as a synonym for *en_morning*, presumably because the idiom *'het krieken van de dag'* translates to *'the crack of dawn'*.

[2] Our approach is not suited for languages for which no lexical resources exist. However, many languages have some coverage in cross-lingual lexicons. For instance, BabelNet 3.7 automatically aligns WordNet to Wikipedia, providing accurate cross-lingual mappings between 271 languages. In our evaluation, we demonstrate substantial gains for Hebrew and Croatian, both of which are spoken by less than 10 million people worldwide.

2 Related Work

2.1 Semantic Specialization

The usefulness of distributional word representations has been demonstrated across many application areas: Part-of-Speech (POS) tagging (Collobert et al., 2011), machine translation (Zou et al., 2013; Devlin et al., 2014), dependency and semantic parsing (Socher et al., 2013a; Bansal et al., 2014; Chen and Manning, 2014; Johannsen et al., 2015; Ammar et al., 2016), sentiment analysis (Socher et al., 2013b), named entity recognition (Turian et al., 2010; Guo et al., 2014), and many others. The importance of semantic specialization for downstream tasks is relatively unexplored, with improvements in performance so far observed for dialogue state tracking (Mrkšić et al., 2016; Mrkšić et al., 2017), spoken language understanding (Kim et al., 2016b; Kim et al., 2016a) and judging lexical entailment (Vulić et al., 2016).

Semantic specialization methods fall (broadly) into two categories: **a)** those which train distributed representations 'from scratch' by combining distributional knowledge and lexical information; and **b)** those which *inject* lexical information into pre-trained collections of word vectors. Methods from both categories make use of similar lexical resources; common examples include WordNet (Miller, 1995), FrameNet (Baker et al., 1998) or the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013).

Learning from Scratch Some methods modify the prior or the regularization of the original training procedure using the set of linguistic constraints (Yu and Dredze, 2014; Xu et al., 2014; Bian et al., 2014; Kiela et al., 2015; Aletras and Stevenson, 2015). Other methods modify the skip-gram (Mikolov et al., 2013b) objective function by introducing semantic constraints (Yih et al., 2012; Liu et al., 2015) to train word vectors which emphasize word similarity over relatedness. Osborne et al. (2016) propose a method for incorporating prior knowledge into the Canonical Correlation Analysis (CCA) method used by Dhillon et al. (2015) to learn spectral word embeddings. While such methods introduce semantic similarity constraints extracted from lexicons, approaches such as the one proposed by Schwartz et al. (2015) use *symmetric patterns* (Davidov and Rappoport, 2006) to push away antonymous words in their pattern-based vector space. Ono et al. (2015) combines both approaches, using thesauri and distributional data to train embeddings specialized for capturing antonymy. Faruqui and Dyer (2015) use many different lexicons to create interpretable sparse binary vectors which achieve competitive performance across a range of intrinsic evaluation tasks.

In theory, word representations produced by models which consider distributional and lexical information jointly could be as good (or better) than representations produced by fine-tuning distributional vectors. However, their performance has not surpassed that of fine-tuning methods.[3]

Fine-Tuning Pre-trained Vectors Rothe and Schütze (2015) fine-tune word vector spaces to improve the representations of synsets/lexemes found in WordNet. Faruqui et al. (2015) and Jauhar et al. (2015) use synonymy constraints in a procedure termed *retrofitting* to bring the vectors of semantically similar words close together, while Wieting et al. (2015) modify the skip-gram objective function to fine-tune word vectors by injecting paraphrasing constraints from PPDB. Mrkšić et al. (2016) build on the retrofitting approach by jointly injecting synonymy and antonymy constraints; the same idea is reassessed by Nguyen et al. (2016). Kim et al. (2016a) further expand this line of work by incorporating semantic intensity information for the constraints, while Recski et al. (2016) use ensembles of rich *concept dictionaries* to further improve a combined collection of semantically specialized word vectors.

ATTRACT-REPEL is an instance of the second family of models, providing a portable, light-weight approach for incorporating external knowledge into arbitrary vector spaces. In our experiments, we show that ATTRACT-REPEL outperforms previously proposed post-processors, setting the new state-of-art performance on the widely used SimLex-999 word similarity dataset. Moreover, we show that starting from distributional vectors allows our method to use existing cross-lingual resources to tie distributional vector spaces of different languages into a unified vector space which benefits from positive semantic transfer between its constituent languages.

[3]The SimLex-999 web page (`www.cl.cam.ac.uk/~fh295/simlex.html`) lists models with state-of-the-art performance, none of which learn representations jointly.

2.2 Cross-Lingual Word Representations

Most existing models which induce cross-lingual word representations rely on cross-lingual distributional information (Klementiev et al., 2012; Zou et al., 2013; Soyer et al., 2015; Huang et al., 2015, inter alia). These models differ in the cross-lingual signal/supervision they use to tie languages into unified bilingual vector spaces: some models learn on the basis of parallel word-aligned data (Luong et al., 2015; Coulmance et al., 2015) or sentence-aligned data (Hermann and Blunsom, 2014a; Hermann and Blunsom, 2014b; Chandar et al., 2014; Gouws et al., 2015). Other models require document-aligned data (Søgaard et al., 2015; Vulić and Moens, 2016), while some learn on the basis of available bilingual dictionaries (Mikolov et al., 2013a; Faruqui and Dyer, 2014; Lazaridou et al., 2015; Vulić and Korhonen, 2016b; Duong et al., 2016). See Upadhyay et al. (2016) and Vulić and Korhonen (2016b) for an overview of cross-lingual word embedding work.

The inclusion of cross-lingual information results in shared cross-lingual vector spaces which can: **a)** boost performance on monolingual tasks such as word similarity (Faruqui and Dyer, 2014; Rastogi et al., 2015; Upadhyay et al., 2016); and **b)** support cross-lingual tasks such as bilingual lexicon induction (Mikolov et al., 2013a; Gouws et al., 2015; Duong et al., 2016), cross-lingual information retrieval (Vulić and Moens, 2015; Mitra et al., 2016), and transfer learning for resource-lean languages (Søgaard et al., 2015; Guo et al., 2015).

However, prior work on cross-lingual word embedding has tended not to exploit pre-existing linguistic resources such as BabelNet. In this work, we make use of cross-lingual constraints derived from such repositories to induce high-quality cross-lingual vector spaces by facilitating semantic transfer from high- to lower-resource languages. In our experiments, we show that cross-lingual vector spaces produced by ATTRACT-REPEL consistently outperform a representative selection of five strong cross-lingual word embedding models in both intrinsic and extrinsic evaluation across several languages.

3 The ATTRACT-REPEL Model

In this section, we propose a new algorithm for producing semantically specialized word vectors by injecting similarity and antonymy constraints into distributional vector spaces. This procedure, which we term ATTRACT-REPEL, builds on the Paragram (Wieting et al., 2015) and counter-fitting procedures (Mrkšić et al., 2016), both of which inject linguistic constraints into existing vector spaces to improve their ability to capture semantic similarity.

Let V be the vocabulary, S the set of synonymous word pairs (e.g. *intelligent* and *brilliant*), and A the set of antonymous word pairs (e.g. *vacant* and *occupied*). The optimization procedure operates over mini-batches of synonym and antonym pairs $\mathcal{B}_S$ and $\mathcal{B}_A$ (which list k_1 synonym and k_2 antonym pairs). For ease of notation, let each word pair (x_l, x_r) in these two sets correspond to a vector pair $(\mathbf{x}_l, \mathbf{x}_r)$, so that a mini-batch is given by $\mathcal{B}_S = [(\mathbf{x}_l^1, \mathbf{x}_r^1), \ldots, (\mathbf{x}_l^{k_1}, \mathbf{x}_r^{k_1})]$ (similarly for $\mathcal{B}_A$).

Next, we define $T_S = [(\mathbf{t}_l^1, \mathbf{t}_r^1), \ldots, (\mathbf{t}_l^{k_1}, \mathbf{t}_r^{k_1})]$ and $T_A = [(\mathbf{t}_l^1, \mathbf{t}_r^1), \ldots, (\mathbf{t}_l^{k_2}, \mathbf{t}_r^{k_2})]$ as pairs of *negative examples* for each synonymy and antonymy example pair in mini-batches $\mathcal{B}_S$ and $\mathcal{B}_A$. These negative examples are chosen from the word vectors present in $\mathcal{B}_S$ or $\mathcal{B}_A$ so that:

- For each synonymy pair $(\mathbf{x}_l, \mathbf{x}_r)$, the negative example pair $(\mathbf{t}_l, \mathbf{t}_r)$ is chosen from the remaining in-batch vectors so that $\mathbf{t}_l$ is the one closest (cosine similarity) to $\mathbf{x}_l$ and $\mathbf{t}_r$ is closest to $\mathbf{x}_r$.
- For each antonymy pair $(\mathbf{x}_l, \mathbf{x}_r)$, the negative example pair $(\mathbf{t}_l, \mathbf{t}_r)$ is chosen from the remaining in-batch vectors so that $\mathbf{t}_l$ is the one furthest away from $\mathbf{x}_l$ and $\mathbf{t}_r$ is the one furthest from $\mathbf{x}_r$.

These negative examples are used to: **a)** force synonymous pairs to be closer to each other than to their respective negative examples; and **b)** to force antonymous pairs to be further away from each other than from their negative examples. The first term of the cost function pulls synonymous words together:

$$S(\mathcal{B}_S, T_S) = \sum_{i=1}^{k_1} [\, \tau\left(\delta_{syn} + \mathbf{x}_l^i \mathbf{t}_l^i - \mathbf{x}_l^i \mathbf{x}_r^i\right) + \tau\left(\delta_{syn} + \mathbf{x}_r^i \mathbf{t}_r^i - \mathbf{x}_l^i \mathbf{x}_r^i\right)]$$

where $\tau(x) = \max(0, x)$ is the hinge loss function and δ_{syn} is the similarity margin which determines how much closer synonymous vectors should be to

each other than to their respective negative examples. The second part of the cost function pushes antonymous word pairs away from each other:

$$A(\mathcal{B}_A, T_A) = \sum_{i=1}^{k_2} [\, \tau\left(\delta_{ant} + \mathbf{x}_l^i \mathbf{x}_r^i - \mathbf{x}_l^i \mathbf{t}_l^i\right) + \tau\left(\delta_{ant} + \mathbf{x}_l^i \mathbf{x}_r^i - \mathbf{x}_r^i \mathbf{t}_r^i\right)]$$

In addition to these two terms, we include an additional regularization term which aims to *preserve* the abundance of high-quality semantic content present in the initial (distributional) vector space, as long as this information does not contradict the injected linguistic constraints. If $V(\mathcal{B})$ is the set of all word vectors present in the given mini-batch, then:

$$R(\mathcal{B}_S, \mathcal{B}_A) = \sum_{\mathbf{x}_i \in V(\mathcal{B}_S \cup \mathcal{B}_A)} \lambda_{reg} \left\| \widehat{\mathbf{x}}_i - \mathbf{x}_i \right\|_2$$

where λ_{reg} is the L2 regularization constant and $\widehat{\mathbf{x}}_i$ denotes the original (distributional) word vector for word x_i. The final ATTRACT-REPEL cost function is given by the sum of all three terms:

$$C(\mathcal{B}_S, T_S, \mathcal{B}_A, T_A) = S(\mathcal{B}_S, T_S) + A(\mathcal{B}_A, T_A) + R(\mathcal{B}_S, \mathcal{B}_A)$$

Comparison to Prior Work ATTRACT-REPEL draws inspiration from three methods: **1)** retrofitting (Faruqui et al., 2015); **2)** PARAGRAM (Wieting et al., 2015); and **3)** counter-fitting (Mrkšić et al., 2016). Whereas retrofitting and PARAGRAM do not consider antonymy, counter-fitting models both synonymy and antonymy. ATTRACT-REPEL differs from this method in two important ways:

1. **Context-Sensitive Updates:** Counter-fitting uses attract and repel terms which pull synonyms together and push antonyms apart without considering their relation to other word vectors. For example, its attract term is given by:

   $$Attract(S) = \sum_{(\mathbf{x}_l, \mathbf{x}_r) \in S} \tau(\delta_{syn} - \mathbf{x}_l \mathbf{x}_r)$$

   where S is the set of synonyms and δ_{syn} is the synonymy margin. Conversely, ATTRACT-REPEL fine-tunes vector spaces by operating over mini-batches of example pairs, updating word vectors only if the position of their negative example implies a stronger semantic relation than that expressed by the position of its target example. Importantly, ATTRACT-REPEL makes fine-grained updates to both the example pair *and* the negative examples, rather than updating the example word pair but ignoring how this affects its relation to all other word vectors.

2. **Regularization:** Counter-fitting preserves distances between pairs of word vectors in the initial vector space, trying to 'pull' the words' neighborhoods with them as they move to incorporate external knowledge. The radius of this initial neighborhood introduces an opaque hyperparameter to the procedure. Conversely, ATTRACT-REPEL implements standard L2 regularization, which 'pulls' each vector towards its distributional vector representation.

In our intrinsic evaluation (Section 5), we perform an exhaustive comparison of these models, showing that ATTRACT-REPEL outperforms counter-fitting in both mono- and cross-lingual setups.

Optimization Following Wieting et al. (2015), we use the AdaGrad algorithm (Duchi et al., 2011) to train the word embeddings for five epochs, which suffices for the parameter estimates to converge. Similar to Faruqui et al. (2015), Wieting et al. (2015) and Mrkšić et al. (2016), we do not use early stopping. By not relying on language-specific validation sets, the ATTRACT-REPEL procedure can induce semantically specialized word vectors for languages with no intrinsic evaluation datasets.[4]

Hyperparameter Tuning We use Spearman's correlation of the final word vectors with the Multilingual WordSim-353 gold-standard association dataset (Finkelstein et al., 2002; Leviant and Reichart, 2015). The ATTRACT-REPEL procedure has six hyperparameters: the regularization constant λ_{reg}, the similarity and antonymy margins δ_{sim} and δ_{ant}, mini-batch sizes k_1 and k_2, and the size of the PPDB constraint set used for each language (larger sizes include more

[4]Many languages are present in semi-automatically constructed lexicons such as BabelNet or PPDB (see the discussion in Section 4.2.). However, intrinsic evaluation datasets such as SimLex-999 exist for very few languages, as they require expert translators and skilled annotators.

	English		German		Italian		Russian	
	syn	ant	syn	ant	syn	ant	syn	ant
English	640	5	246	11	356	24	196	9
German	-	-	135	2	277	13	175	6
Italian	-	-	-	-	159	7	220	11
Russian	-	-	-	-	-	-	48	1

Table 2: Linguistic constraint counts (in thousands). For each language pair, the two figures show the number of injected synonymy and antonymy constraints. Monolingual constraints (the diagonal elements) are underlined.

constraints, but also a larger proportion of false synonyms). We ran a grid search over these for the four SimLex languages, choosing the hyperparameters which achieved the best WordSim-353 score.[5]

4 Experimental Setup

4.1 Distributional Vectors

We first present our sixteen experimental languages: English (EN), German (DE), Italian (IT), Russian (RU), Dutch (NL), Swedish (SV), French (FR), Spanish (ES), Portuguese (PT), Polish (PL), Bulgarian (BG), Croatian (HR), Irish (GA), Persian (FA) and Vietnamese (VI). The first four languages are those of the Multilingual SimLex-999 dataset.

For the four SimLex languages, we employ four well-known, high-quality word vector collections: **a)** The Common Crawl GloVe English vectors from Pennington et al. (2014); **b)** German vectors from Vulić and Korhonen (2016a); **c)** Italian vectors from Dinu et al. (2015); and **d)** Russian vectors from Kutuzov and Andreev (2015). In addition, for each of the 16 languages we also train the skip-gram with negative sampling variant of the `word2vec` model (Mikolov et al., 2013b), on the latest Wikipedia dump of each language, to induce 300-dimensional word vectors.[6]

[5]We ran the grid search over $\lambda_{reg} \in [10^{-3}, \ldots, 10^{-10}]$, $\delta_{sim}, \delta_{ant} \in [0, 0.1, \ldots, 1.0]$, $k_1, k_2 \in [10, 25, 50, 100, 200]$ and over the six PPDB sizes for the four SimLex languages. $\lambda_{reg} = 10^{-9}$, $\delta_{sim} = 0.6$, $\delta_{ant} = 0.0$ and $k_1 = k_2 \in [10, 25, 50]$ consistently achieved the best performance (we use $k_1 = k_2 = 50$ in all experiments for consistency). The PPDB constraint set size *XL* was best for English, German and Italian, and *M* achieved the best performance for Russian.

[6]The frequency cut-off was set to 50: words that occurred less frequently were removed from the vocabularies. Other `word2vec` parameters were set to the standard values (Vulić and Korhonen, 2016a): 15 epochs, 15 negative samples, global (decreasing) learning rate: 0.025, subsampling rate: $1e-4$.

4.2 Linguistic Constraints

Table 2 shows the number of monolingual and cross-lingual constraints for the four SimLex languages.

Monolingual Similarity We employ the Multilingual Paraphrase Database (Ganitkevitch and Callison-Burch, 2014). This resource contains paraphrases automatically extracted from parallel-aligned corpora for ten of our sixteen languages. In our experiments, the remaining six languages (HE, HR, SV, GA, VI, FA) serve as examples of *lower-resource* languages, as they have no monolingual synonymy constraints.

Cross-Lingual Similarity We employ BabelNet, a multilingual semantic network automatically constructed by linking Wikipedia to WordNet (Navigli and Ponzetto, 2012; Ehrmann et al., 2014). BabelNet groups words from different languages into *Babel synsets*. We consider two words from any (distinct) language pair to be synonymous if they belong to (at least) one set of synonymous Babel synsets. We made use of all BabelNet word senses tagged as *conceptual* but ignored the ones tagged as *Named Entities*.

Given a large collection of *cross-lingual* semantic constraints (e.g. the translation pair *en_sweet* and *it_dolce*), ATTRACT-REPEL can use them to bring the vector spaces of different languages together into a shared cross-lingual space. Ideally, sharing information across languages should lead to improved semantic content for each language, especially for those with limited monolingual resources.

Antonymy BabelNet is also used to extract *both* monolingual *and* cross-lingual antonymy constraints. Following Faruqui et al. (2015), who found PPDB constraints more beneficial than the WordNet ones, we do not use BabelNet for monolingual synonymy.

Availability of Resources Both PPDB and BabelNet are created automatically. However, PPDB relies on large, high-quality parallel corpora such as Europarl (Koehn, 2005). In total, Multilingual PPDB provides collections of paraphrases for 22 languages. On the other hand, BabelNet uses Wikipedia's *interlanguage links* and statistical machine translation (Google Translate) to provide cross-lingual mappings for 271 languages. In our evaluation, we show that PPDB and BabelNet can be used jointly to improve word representations for lower-resource languages by

tying them into bilingual spaces with high-resource ones. We validate this claim on Hebrew and Croatian, which act as 'lower-resource' languages because of their lack of any PPDB resource and their relatively small Wikipedia sizes.[7]

5 Intrinsic Evaluation

5.1 Datasets

Spearman's rank correlation with the SimLex-999 dataset (Hill et al., 2015) is used as the intrinsic evaluation metric throughout the experiments. Unlike other gold standard resources such as WordSim-353 (Finkelstein et al., 2002) or MEN (Bruni et al., 2014), SimLex-999 consists of word pairs scored by annotators instructed to discern between semantic similarity and conceptual association, so that related but non-similar words (e.g. *book* and *read*) have a low rating.

Leviant and Reichart (2015) translated SimLex-999 to German, Italian and Russian, crowd-sourcing the similarity scores from native speakers of these languages. We use this resource for multilingual intrinsic evaluation.[8] To investigate the portability of our approach to lower-resource languages, we used the same experimental setup to collect SimLex-999 datasets for Hebrew and Croatian.[9] For English vectors, we also report Spearman's correlation with SimVerb-3500 (Gerz et al., 2016), a semantic similarity dataset that focuses on verb pair similarity.

5.2 Experiments

Monolingual and Cross-Lingual Specialization We start from distributional vectors for the SimLex languages: English, German, Italian and Russian. For each language, we first perform semantic specialization of these spaces using: a) monolingual synonyms; b) monolingual antonyms; and c) the combination of both. We then add cross-lingual synonyms and antonyms to these constraints and train a shared four-lingual vector space for these languages.

Comparison to Baseline Methods Both mono- and cross-lingual specialization was performed using ATTRACT-REPEL and counter-fitting, in order to conclusively determine which of the two methods exhibited superior performance. Retrofitting and PARAGRAM methods only inject synonymy, and their cost functions can be expressed using sub-components of counter-fitting and ATTRACT-REPEL cost functions. As such, the performance of the two investigated methods when they make use of synonymy (but not antonymy) constraints illustrates the performance range of the two preceding models.

Importance of Initial Vectors We use three different sets of initial word vectors: **a)** well-known distributional word vector collections (Section 4.1); **b)** distributional word vectors trained on the latest Wikipedia dumps; and **c)** word vectors randomly initialized using the XAVIER initialization (Glorot and Bengio, 2010).

Specialization for Lower-Resource Languages In this experiment, we first construct bilingual spaces which combine: **a)** one of the four SimLex languages; with **b)** each of the other twelve languages.[10] Since each pair contains at least one SimLex language, we can analyse the improvement over monolingual specialization to understand how robust the performance gains are across different language pairs. We next use the newly collected SimLex datasets for Hebrew and Croatian to evaluate the extent to which bilingual semantic specialization using ATTRACT-REPEL and BabelNet constraints can improve word representations for lower-resource languages.

Comparison to State-of-the-Art Bilingual Spaces The English-Italian and English-German bilingual spaces induced by ATTRACT-REPEL were compared to five state-of-the-art methods for constructing bilingual vector spaces: **1.** (Mikolov et al., 2013a), retrained using the constraints used by our model; and **2.-5.** (Hermann and Blunsom, 2014a; Gouws et al., 2015; Vulić and Korhonen, 2016a; Vulić and Moens, 2016). The latter models use various sources of supervision (word-, sentence- and document-aligned

[7] Hebrew and Croatian Wikipedias (which are used to induce their BabelNet constraints) currently consist of 203,867 / 172,824 articles, ranking them 40th / 42nd by size.

[8] Leviant and Reichart (2015) also re-scored the original English SimLex. We report results on their version, but also provide numbers for the original dataset for comparability.

[9] The 999 word pairs and annotator instructions were translated by native speakers and scored by 10 annotators. The inter-annotator agreement scores (Spearman's ρ) were 0.77 (pairwise) and 0.87 (mean) for Croatian, and 0.59 / 0.71 for Hebrew.

[10] Hyperparameters: we used $\delta_{sim} = 0.6$, $\delta_{ant} = 0.0$ and $\lambda_{reg} = 10^{-9}$, which achieved the best performance when tuned for the original SimLex languages. The largest available PPDB size was used for the six languages with available PPDB (French, Spanish, Portuguese, Polish, Bulgarian and Dutch).

Word Vectors	English	German	Italian	Russian
Monolingual Distributional Vectors	0.32	0.28	0.36	0.38
COUNTER-FITTING: Mono-Syn	0.45	0.24	0.29	0.46
COUNTER-FITTING: Mono-Ant	0.33	0.28	0.47	0.42
COUNTER-FITTING: Mono-Syn + Mono-Ant	0.50	0.26	0.35	0.49
COUNTER-FITTING: Cross-Syn	0.46	0.43	0.45	0.37
COUNTER-FITTING: Mono-Syn + Cross-Syn	0.47	0.40	0.43	0.45
COUNTER-FITTING: Mono-Syn + Mono-Ant + Cross-Syn + Cross-Ant	0.53	0.41	0.49	0.48
ATTRACT-REPEL: Mono-Syn	0.56	0.40	0.46	0.53
ATTRACT-REPEL: Mono-Ant	0.42	0.30	0.45	0.41
ATTRACT-REPEL: Mono-Syn + Mono-Ant	0.65	0.43	0.56	0.56
ATTRACT-REPEL: Cross-Syn	0.57	0.53	0.58	0.46
ATTRACT-REPEL: Mono-Syn + Cross-Syn	0.61	0.58	0.59	0.54
ATTRACT-REPEL: Mono-Syn + Mono-Ant + Cross-Syn + Cross-Ant	**0.71**	**0.62**	**0.67**	**0.61**

Table 3: Multilingual SimLex-999. The effect of using the COUNTER-FITTING and ATTRACT-REPEL procedures to inject mono- and cross-lingual synonymy and antonymy constraints into the four collections of distributional word vectors. Our best results set the new state-of-the-art performance for all four languages.

Word Vectors	EN	DE	IT	RU
Random Init. (No Info.)	0.01	-0.03	0.02	-0.03
A-R: Monolingual Cons.	0.54	0.33	0.29	0.35
A-R: Mono + Cross-Ling.	0.66	0.49	0.59	0.51
Distributional Wiki Vectors	0.32	0.31	0.28	0.19
A-R: Monolingual Cons.	0.61	0.48	0.53	0.52
A-R: Mono + Cross-Ling.	0.66	0.60	0.65	0.54

Table 4: Multilingual SimLex-999. The effect of ATTRACT-REPEL (A-R) on alternative sets of starting word vectors (Random = XAVIER initialization).

corpora), which means they cannot be trained using our sets of constraints. For these models, we use competitive setups proposed in (Vulić and Korhonen, 2016a). The goal of this experiment is to show that vector spaces induced by ATTRACT-REPEL exhibit better intrinsic and extrinsic performance when deployed in language understanding tasks.

5.3 Results and Discussion

Table 3 shows the effects of monolingual and cross-lingual semantic specialization of four well-known distributional vector spaces for the SimLex languages. Monolingual specialization leads to very strong improvements in the SimLex performance across all languages. Cross-lingual specialization brings further improvements, with all languages benefiting from sharing the cross-lingual vector space. German and Italian in particular show strong evidence of effective transfer (+0.19 / +0.11 over monolingual specialization), with Italian vectors' performance coming close to the top-performing English ones.

Comparison to Baselines Table 3 gives an exhaustive comparison of ATTRACT-REPEL to counter-fitting: ATTRACT-REPEL achieved substantially stronger performance in all experiments. We believe these results conclusively show that the context-sensitive updates and L2 regularization employed by ATTRACT-REPEL present a better alternative to the context-insensitive attract/repel terms and pair-wise regularization employed by counter-fitting.[11]

State-of-the-Art Wieting et al. (2016) note that the hyperparameters of the widely used Paragram-SL999 vectors (Wieting et al., 2015) are tuned on SimLex-999, and as such are not comparable to methods which hold out the dataset. This implies that further work which uses these vectors (e.g., (Mrkšić et al.,

[11]To understand the relative importance of the context-sensitive updates and the change in regularization, we can compare the two methods to the retrofitting procedure (Faruqui et al., 2015). Retrofitting uses L2 regularization (like ATTRACT-REPEL) and a 'global' attract term (like counter-fitting). The performance of retrofitting using all mono- and cross-lingual synonymy constraints (the procedure does not support antonyms) gives an EN-DE-IT-RU score of [0.41, 0.30, 0.36, 0.40], which is a change of [-0.06, -0.10, -0.07, -0.08] compared to counter-fitting, and is substantially weaker than ATTRACT-REPEL: [-0.20, -0.28, -0.23, -0.21]. We can therefore conclude that the bulk of the performance improvement achieved by ATTRACT-REPEL stems from using the context-sensitive updates. Counter-fitting outperforms retrofitting as well, which implies that its pairwise regularization is an improvement over simple L2 regularization. However, its quadratic complexity makes it intractable for the scale of experiments performed in this paper (unlike ATTRACT-REPEL, which fine-tunes the vector space in less than 2 minutes using an NVIDIA GeForce GTX 1080 graphics card).

	Mono.	SimLex Languages				PPDB available						No PPDB available					
	Spec.	**EN**	**DE**	**IT**	**RU**	**NL**	**FR**	**ES**	**PT**	**PL**	**BG**	**HR**	**HE**	**GA**	**VI**	**FA**	**SV**
English	0.65	-	0.69	0.70	0.70	0.70	0.72	0.72	0.70	0.70	0.68	0.70	0.66	0.65	0.67	0.68	0.70
German	0.43	0.61	-	0.58	0.56	0.55	0.60	0.59	0.56	0.54	0.52	0.53	0.50	0.49	0.48	0.51	0.55
Italian	0.56	0.69	0.65	-	0.64	0.67	0.68	0.68	0.66	0.66	0.62	0.63	0.59	0.60	0.58	0.61	0.63
Russian	0.56	0.63	0.59	0.62	-	0.61	0.61	0.62	0.58	0.60	0.61	0.59	0.56	0.57	0.58	0.58	0.60

Table 5: SimLex-999 performance. Tying the SimLex languages into bilingual vector spaces with 16 different languages. The first number in each row represents monolingual specialization. All but two of the bilingual spaces improved over these baselines. The EN-FR vectors set a new high score of **0.754** on the original (English) SimLex-999.

2016; Recski et al., 2016)) as a starting point does not yield meaningful high scores either. Our reported English score of 0.71 on the Multilingual SimLex-999 corresponds to **0.751** on the original SimLex-999: it outperforms the 0.706 score reported by Wieting et al. (2016) and sets a new high score for this dataset. Similarly, the SimVerb-3500 score of these vectors is **0.674**, outperforming the current state-of-the-art score of 0.628 reported by Gerz et al. (2016).

Starting Distributional Spaces Table 4 repeats the previous experiment with two different sets of initial vector spaces: **a)** randomly initialized word vectors;[12] and **b)** skip-gram with negative sampling vectors trained on the latest Wikipedia dumps. The randomly initialized vectors serve to decouple the impact of injecting external knowledge from the information embedded in the distributional vectors. The random vectors benefit from both mono- and cross-lingual specialization: the English performance is surprisingly strong, with other languages suffering more from the lack of initialization.

When comparing distributional vectors trained on Wikipedia to the high-quality word vector collections used in Table 3, the Italian and Russian vectors in particular start from substantially weaker SimLex scores. The difference in performance is largely mitigated through semantic specialization. However, all vector spaces still exhibit a weaker performance compared to those in Table 3. We believe this shows that the quality of the initial distributional vector spaces is important, but can, in large part, be compensated for through semantic specialization.

Bilingual Specialization Table 5 shows the effect of combining the four original SimLex languages with each other and with twelve other languages (Section 4.1). Bilingual specialization substantially improves over monolingual specialization for *all language pairs*. This indicates that our improvements are language independent to a large extent.

Interestingly, even though we use no monolingual synonymy constraints for the six right-most languages, combining them with the SimLex languages still improved word vector quality for these four high-resource languages. The reason why even resource-deprived languages such as Irish help improve vector space quality of high-resource languages such as English or Italian is that they provide implicit indicators of semantic similarity. English words which map to the same Irish word are likely to be synonyms, even if those English pairs are not present in the PPDB datasets (Faruqui and Dyer, 2014).[13]

Lower-Resource Languages The previous experiment indicates that bilingual specialization further improves the (already) high-quality estimates for high-resource languages. However, it does little to show how much (or if) the word vectors of lower-resource languages improve during such specialization. Table 6 investigates this proposition using the newly collected SimLex datasets for Hebrew and Croatian.

Tying the distributional vectors for these languages (which have no monolingual constraints) into cross-lingual spaces with high-resource ones (which do, in our case from PPDB) leads to substantial improvements. Table 6 also shows how the distributional vectors of the four SimLex languages improve when tied to other languages (in each row, we use monolingual constraints only for the 'added' lan-

[12]The XAVIER initialization populates the values for each word vector by uniformly sampling from the interval $[-\frac{\sqrt{6}}{\sqrt{d}}, +\frac{\sqrt{6}}{\sqrt{d}}]$, where d is the vector dimensionality. This is a typical init method in neural nets research (Goldberg, 2015; Bengio et al., 2013).

[13]We release bilingual vector spaces for EN + 51 other languages: the 16 presented here and another 35 languages (all available at `www.github.com/nmrksic/attract-repel`).

	Distrib.	+ EN	+ DE	+ IT	+ RU
Hebrew	0.28	**0.51**	0.46	0.52	0.45
Croatian	0.21	**0.62**	0.49	0.58	0.54
English	0.32	-	0.61	**0.66**	0.63
German	0.28	**0.58**	-	0.55	0.49
Italian	0.36	**0.69**	0.66	-	0.63
Russian	0.38	**0.56**	0.52	0.55	-

Table 6: Bilingual semantic specialization for: **a)** Hebrew and Croatian; and **b)** the original SimLex languages. Each row shows how SimLex scores for that language improve when its distributional vectors are tied into bilingual vector spaces with the four high-resource languages.

guage). Hebrew and Croatian exhibit similar trends to the original SimLex languages: tying to English and Italian leads to stronger gains than tying to the morphologically sophisticated German and Russian. Indeed, tying to English consistently led to the strongest performance. We believe this shows that bilingual ATTRACT-REPEL specialization with English promises to produce high-quality vector spaces for many lower-resource languages which have coverage among the 271 BabelNet languages (but are not available in PPDB).

Existing Bilingual Spaces Table 7 compares the intrinsic (i.e. SimLex-999) performance of bilingual English-Italian and English-German vectors produced by ATTRACT-REPEL to five previously proposed approaches for constructing bilingual vector spaces. For both languages in both language pairs, ATTRACT-REPEL achieves substantial gains over all of these methods. In the next section, we show that these differences in intrinsic performance lead to substantial gains in downstream evaluation.

Model	EN-IT		EN-DE	
	EN	IT	EN	DE
(Mikolov et al., 2013a)	0.32	0.28	0.32	0.28
(Hermann and Blunsom, 2014a)	0.40	0.34	0.38	0.35
(Gouws et al., 2015)	0.25	0.18	0.25	0.14
(Vulić and Korhonen, 2016a)	0.32	0.27	0.32	0.33
(Vulić and Moens, 2016)	0.23	0.25	0.20	0.25
Bilingual ATTRACT-REPEL	**0.70**	**0.69**	**0.69**	**0.61**

Table 7: Comparison of the intrinsic quality (SimLex-999) of bilingual spaces produced by the ATTRACT-REPEL method to those produced by five state-of-the-art methods for constructing bilingual vector spaces.

User: Suggest something fancy.
`[price=expensive]`
System: Sure, where?
User: Downtown. Any Korean places?
`[price=expensive, area=centre, food=Korean]`
System: Sorry, no Korean places in the centre.
User: How about Japanese?
`[price=expensive, area=centre, food=Japanese]`
System: Sticks'n'Sushi meets your criteria.

Figure 1: Annotated dialogue states in a sample dialogue. Underlined words show rephrasings for ontology values which are typically handled using semantic dictionaries.

6 Downstream Task Evaluation

6.1 Dialogue State Tracking

Task-oriented dialogue systems help users achieve goals such as making travel reservations or finding restaurants. In *slot-based* systems, application domains are defined by *ontologies* which enumerate the goals that users can express (Young, 2010). The goals are expressed by *slot-value* pairs such as [price: *cheap*] or [food: *Thai*]. For modular task-based systems, the Dialogue State Tracking (DST) component is in charge of maintaining the *belief state*, which is the system's internal distribution over the possible states of the dialogue. Figure 1 shows the correct dialogue state for each turn of an example dialogue.

Unseen Data/Labels As dialogue ontologies can be very large, many of the possible *class labels* (i.e., the various *food types* or *street names*) will not occur in the training set. To overcome this problem, *delexicalization-based* DST models (Henderson et al., 2014c; Henderson et al., 2014b; Mrkšić et al., 2015; Wen et al., 2017) replace occurrences of ontology values with generic tags which facilitate transfer learning across different ontology values. This is done through exact matching supplemented with *semantic lexicons* which encode rephrasings, morphology and other linguistic variation. For instance, such lexicons would be required to deal with the underlined non-exact matches in Figure 1.

Exact Matching as a Bottleneck Semantic lexicons can be hand-crafted for small dialogue domains.

Mrkšić et al. (2016) showed that semantically specialized vector spaces can be used to automatically induce such lexicons for simple dialogue domains. However, as domains grow more sophisticated, the reliance on (manually- or automatically-constructed) semantic dictionaries which list potential rephrasings for ontology values becomes a bottleneck for deploying dialogue systems. Ambiguous rephrasings are just one problematic instance of this approach: a user asking about *Iceland* could be referring to the country or the supermarket chain, and someone asking for songs by *Train* is not interested in train timetables. More importantly, the use of English as the principal language in most dialogue systems research understates the challenges that complex linguistic phenomena present in other languages. In this work, we investigate the extent to which semantic specialization can empower DST models which *do not rely* on such dictionaries.

Neural Belief Tracker (NBT) The NBT is a novel DST model which operates purely over distributed representations of words, learning to compose utterance and context representations which it then uses to decide which of the potentially many ontology-defined intents (goals) have been expressed by the user (Mrkšić et al., 2017). To overcome the data sparsity problem, the NBT uses *label embedding* to decompose this multi-class classification problem into many binary classification ones: for each slot, the model iterates over slot values defined by the ontology, deciding whether each of them was expressed in the current utterance and its surrounding context. The first NBT layer consists of neural networks which produce distributed representations of the user utterance,[14] the preceding system output and the *embedded label* of the candidate slot-value pair. These representations are then passed to the downstream *semantic decoding* and *context modelling* networks, which subsequently make the binary decision regarding the current slot-value candidate. When contradicting goals are detected (i.e. *cheap* and *expensive*), the model chooses the more probable one.

The NBT training procedure keeps the initial word vectors fixed. That way, at test time, unseen words semantically related to familiar slot values (i.e. *affordable* or *cheaper* to *cheap*) are recognized purely by their position in the original vector space. Thus, it is essential that deployed word vectors are specialized for semantic similarity, as distributional effects which keep antonymous words' vectors together can be very detrimental to DST performance (e.g., by matching *northern* to *south* or *inexpensive* to *expensive*).

The Multilingual WOZ 2.0 Dataset Our DST evaluation is based on the WOZ 2.0 dataset introduced by Wen et al. (2017) and Mrkšić et al. (2017). This dataset is based on the ontology used for the 2nd DST Challenge (DSTC2) (Henderson et al., 2014a). It consists of 1,200 Wizard-of-Oz (Fraser and Gilbert, 1991) dialogues in which Amazon Mechanical Turk users assumed the role of the dialogue system or the caller looking for restaurants in Cambridge, UK. Since users typed instead of using speech and interacted with intelligent assistants, the language they used was more sophisticated than in case of DSTC2, where users would quickly adapt to the system's inability to cope with complex queries. For our experiments, the ontology and 1,200 dialogues were translated to Italian and German through `gengo.com`, a web-based human translation platform.

6.2 DST Experiments

The principal evaluation metric in our DST experiments is the *joint goal accuracy*, which represents the proportion of test set dialogue turns where all the search constraints expressed up to that point in the conversation were decoded correctly. Our DST experiments investigate two propositions:

1. Intrinsic vs. Downstream Evaluation If mono- and cross-lingual semantic specialization improves the semantic content of word vector collections according to intrinsic evaluation, we would expect the NBT model to perform higher-quality belief tracking when such improved vectors are deployed. We investigate the difference in DST performance for English, German and Italian when the NBT model employs the following word vector collections: **1)** distributional word vectors; **2)** monolingual semantically specialized vectors; and **3)** monolingual subspaces of the cross-lingual semantically specialized EN-DE-IT-RU vectors. For each language, we also compare to the NBT performance achieved using

[14] There are two variants of the NBT model: **NBT-DNN** and **NBT-CNN**. In this work, we limit our investigation to the latter one, as it achieved consistently stronger DST performance.

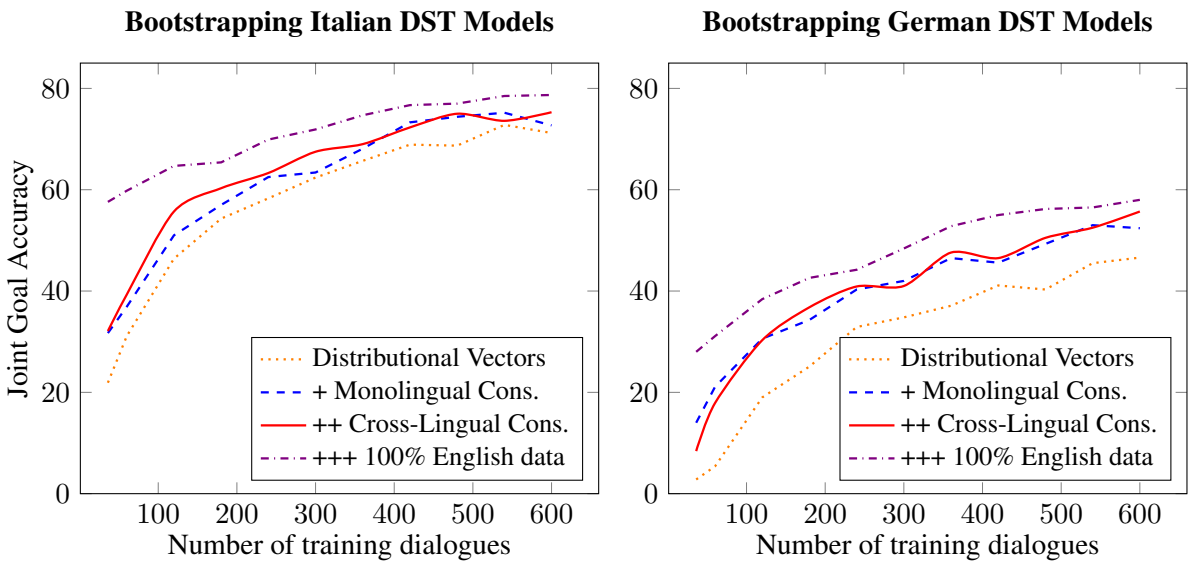

Figure 2: Joint goal accuracy of the NBT-CNN model for Italian (left) and German (right) WOZ 2.0 test sets as a function of the number of in-language dialogues used for training.

the five state-of-the-art bilingual vector spaces we compared to in Section 5.3.

2. Training a Multilingual DST Model The values expressed by the domain ontology (e.g., *cheap*, *north*, *Thai*, etc.) are language independent. If we assume common semantic grounding across languages, we can *decouple* the ontologies from the dialogue corpora and use a *single ontology* (i.e. its values' vector representations) across all languages. Since we know that high-performing English DST is attainable, we will *ground* the Italian and German ontologies (i.e. all slot-value pairs) to the original English ontology. The use of a single ontology coupled with cross-lingual vectors then allows us to combine the training data for multiple languages and train a single NBT model capable of performing belief tracking across all three languages at once. Given a high-quality cross-lingual vector space, combining the languages effectively increases the training set size and should therefore lead to improved performance across all languages.

6.3 Results and Discussion

The DST performance of the NBT-CNN model on English, German and Italian WOZ 2.0 datasets is shown in Table 8. The first five rows show the performance when the model employs the five baseline vector spaces. The subsequent three rows show the performance of: **a)** distributional vector spaces; **b)** their monolingual specialization; and **c)** their EN-DE-IT-RU cross-lingual specialization. The last row shows the performance of the multilingual DST model trained using *ontology grounding*, where the training data of all three languages was combined and used to train an improved model. Figure 2 investigates the usefulness of ontology grounding for bootstrapping DST models for new languages with less data. The two figures display the Italian / German performance of models trained using different proportions of the in-language training dataset. The top-performing dash-dotted curve shows the performance of the model trained using the language-specific dialogues and all of the English training data.

Word Vector Space	EN	IT	DE
EN-IT/EN-DE (Mikolov et al., 2013a)	78.2	71.1	50.5
EN-IT/EN-DE (Hermann et al., 2014a)	71.7	69.3	44.7
EN-IT/EN-DE (Gouws et al., 2015)	75.0	68.4	45.4
EN-IT/EN-DE (Vulić et al., 2016a)	81.6	71.8	50.5
EN-IT/EN-DE (Vulić et al., 2016)	72.3	69.0	38.2
Monolingual Distributional Vectors	77.6	71.2	46.6
A-R: Monolingual Specialization	80.9	72.7	52.4
A-R: Cross-Lingual Specialization	80.3	75.3	55.7
+ English Ontology Grounding	**82.8**	**77.1**	**57.7**

Table 8: NBT model accuracy across the three languages. Each figure shows the performance of the model trained using the subspace of the given vector space corresponding to the target language. For the English baseline figures, we show the stronger of the EN-IT / EN-DE figures.

The results in Table 8 show that both types of specialization improve over DST performance achieved using the distributional vectors or the five baseline bilingual spaces. Interestingly, the bilingual vectors of Vulić and Korhonen (2016a) outperform ours for EN (but not for IT and DE) despite their weaker SimLex performance, showing that intrinsic evaluation does not capture all relevant aspects pertaining to word vectors' usability for downstream tasks.

The multilingual DST model trained using ontology grounding offers substantial performance improvements, with particularly large gains in the low-data scenario investigated in Figure 2 (dash-dotted purple line). This figure also shows that the difference in performance between our mono- and cross-lingual vectors is not very substantial. Again, the large disparity in SimLex scores induced only minor improvements in DST performance.

In summary, our results show that: **a)** semantically specialized vectors benefit DST performance; **b)** large gains in SimLex scores do not always induce large downstream gains; and **c)** high-quality cross-lingual spaces facilitate transfer learning between languages and offer an effective method for bootstrapping DST models for lower-resource languages.

Finally, German DST performance is substantially weaker than both English and Italian, corroborating our intuition that linguistic phenomena such as cases and compounding make German DST very challenging. We release these datasets in the hope that multilingual DST evaluation can give the NLP community a tool for evaluating downstream performance of vector spaces for morphologically richer languages.

7 Conclusion

We have presented a novel ATTRACT-REPEL method for injecting linguistic constraints into word vector space representations. The procedure *semantically specializes* word vectors by jointly injecting mono- *and* cross-lingual synonymy and antonymy constraints, creating unified cross-lingual vector spaces which achieve the state-of-the-art performance on the well-established SimLex-999 dataset and its multilingual variants. Next, we have shown that ATTRACT-REPEL can induce high-quality vectors for lower-resource languages by tying them into bilingual vector spaces with high-resource ones. We also demonstrated that the substantial gains in intrinsic evaluation translate to gains in the downstream task of dialogue state tracking (DST), for which we release two novel non-English datasets (in German and Italian). Finally, we have shown that our semantically rich cross-lingual vectors facilitate language transfer in DST, providing an effective method for bootstrapping belief tracking models for new languages.

Further Work Our results, especially with DST, emphasize the need for improving vector space models for morphologically rich languages. Moreover, our intrinsic and task-based experiments exposed the discrepancies between the conclusions that can be drawn from these two types of evaluation. We consider these to be major directions for future work.

Acknowledgements

The authors would like to thank Anders Johannsen for his help with extracting BabelNet constraints. We would also like to thank our action editor Sebastian Padó and the anonymous TACL reviewers for their constructive feedback. Ivan Vulić, Roi Reichart and Anna Korhonen are supported by the ERC Consolidator Grant LEXICAL (number 648909). Roi Reichart is also supported by the Intel-ICRI grant: Hybrid Models for Minimally Supervised Information Extraction from Conversations.

References

Nikolaos Aletras and Mark Stevenson. 2015. A hybrid distributional and knowledge-based model of lexical semantics. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics, *SEM*, pages 20–29.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. *Transactions of the ACL*, 4:431–444.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of ACL*, pages 86–90.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of ACL*, pages 809–815.

Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

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

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

Sarath A.P. Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh M. Khapra, Balaraman Ravindran, Vikas C. Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Proceedings of NIPS*, pages 1853–1861.

Danqi Chen and Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP*, pages 740–750.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, fast cross-lingual word embeddings. In *Proceedings of EMNLP*, pages 1109–1113.

Dmitry Davidov and Ari Rappoport. 2006. Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. In *Proceedings of ACL*, pages 297–304.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard M. Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of ACL*, pages 1370–1380.

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

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of ICLR: Workshop Papers*.

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.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of EMNLP*, pages 1285–1295.

Maud Ehrmann, Francesco Cecconi, Daniele Vannella, John Philip Mccrae, Philipp Cimiano, and Roberto Navigli. 2014. Representing multilingual data as linked data: The case of BabelNet 2.0. In *Proceedings of LREC*, pages 401–408.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of EACL*, pages 462–471.

Manaal Faruqui and Chris Dyer. 2015. Non-distributional word vector representations. In *Proceedings of ACL*, pages 464–469.

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

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.

Norman M. Fraser and G. Nigel Gilbert. 1991. Simulating speech systems. *Computer Speech and Language*, 5(1):81–99.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The Multilingual Paraphrase Database. In *Proceedings of LREC*, pages 4276–4283.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of NAACL*, pages 758–764.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of EMNLP*, pages 2173–2182.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of AISTATS*, pages 249–256.

Yoav Goldberg. 2015. A primer on neural network models for natural language processing. *CoRR*, abs/1510.00726.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of ICML*, pages 748–756.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Revisiting embedding features for simple semi-supervised learning. In *Proceedings of EMNLP*, pages 110–120.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of ACL*, pages 1234–1244.

Matthew Henderson, Blaise Thomson, and Jason D. Wiliams. 2014a. The Second Dialog State Tracking Challenge. In *Proceedings of SIGDIAL*, pages 263–272.

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

Matthew Henderson, Blaise Thomson, and Steve Young. 2014c. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of SIGDIAL*, pages 292–299.

Karl Moritz Hermann and Phil Blunsom. 2014a. Multilingual Distributed Representations without Word Alignment. In *Proceedings of ICLR*.

Karl Moritz Hermann and Phil Blunsom. 2014b. Multilingual models for compositional distributed semantics. In *Proceedings of ACL*, pages 58–68.

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

Kejun Huang, Matt Gardner, Evangelos Papalexakis, Christos Faloutsos, Nikos Sidiropoulos, Tom Mitchell, Partha P. Talukdar, and Xiao Fu. 2015. Translation invariant word embeddings. In *Proceedings of EMNLP*, pages 1084–1088.

Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. 2014. A Neural Network for Factoid Question Answering over Paragraphs. In *Proceedings of EMLNP*, pages 633–644.

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.

Anders Johannsen, Héctor Martínez Alonso, and Anders Søgaard. 2015. Any-language frame-semantic parsing. In *Proceedings of EMNLP*, pages 2062–2066.

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*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings COLING*, pages 1459–1474.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit, volume 5*.

Andrey Kutuzov and Igor Andreev. 2015. Texts in, meaning out: neural language models in semantic similarity task for Russian. In *Proceedings of DIALOG*.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of ACL*, pages 270–280.

Ira Leviant and Roi Reichart. 2015. Separated by an Un-common Language: Towards Judgment Language Informed Vector Space Modeling. *arXiv preprint: 1508.00106*.

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.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for NLP*, pages 151–159.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint, CoRR*, abs/1309.4168.

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

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, pages 39–41.

Bhaskar Mitra, Eric T. Nalisnick, Nick Craswell, and Rich Caruana. 2016. A dual embedding space model for document ranking. *CoRR*, abs/1602.01137.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2015. Multi-domain dialog state tracking using recurrent neural networks. In *Proceedings of ACL*, pages 794–799.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina 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*, pages 142–148.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Tsung-Hsien Wen, and Steve Young. 2017. Neural Belief Tracker: Data-driven dialogue state tracking. In *Proceedings of ACL*.

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.

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.

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.

Diarmuid Ó Séaghdha and Anna Korhonen. 2014. Probabilistic distributional semantics. *Computational Linguistics*, 40(3):587–631.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Pushpendre Rastogi, Benjamin Van Durme, and Raman Arora. 2015. Multiview LSA: Representation learning via generalized CCA. In *Proceedings of NAACL*, pages 556–566.

Gábor Recski, Eszter Iklódi, Katalin Pajkossy, and Andras Kornai. 2016. Measuring Semantic Similarity of Words Using Concept Networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 193–200.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, and Phil Blunsom. 2016. Reasoning about Entailment with Neural Attention. In *Proceedings of ICLR*.

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

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.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013a. Parsing with compositional vector grammars. In *Proceedings of ACL*, pages 455–465.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013b. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, pages 1631–1642.

Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 2015. Inverted indexing for cross-lingual NLP. In *Proceedings ACL*, pages 1713–1722.

Hubert Soyer, Pontus Stenetorp, and Akiko Aizawa. 2015. Leveraging monolingual data for crosslingual compositional word representations. In *Proceedings of ICLR*.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL*, pages 384–394.

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 Vulić and Anna Korhonen. 2016a. Is "universal syntax" universally useful for learning distributed representations? In *Proceedings of ACL*, pages 518–524.

Ivan Vulić and Anna Korhonen. 2016b. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of ACL*, pages 247–257.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of SIGIR*, pages 363–372.

Ivan Vulić and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2016. Hyperlex: A large-scale evaluation of graded lexical entailment. *CoRR*, abs/1608.02117.

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*, pages 437–449.

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.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Charagram: Embedding words and sentences via character n-grams. In *Proceedings of EMNLP*, pages 1504–1515.

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.

Wen-Tau Yih, Geoffrey Zweig, and John C. Platt. 2012. Polarity inducing Latent Semantic Analysis. In *Proceedings of ACL*, pages 1212–1222.

Steve J. Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. POMDP-Based Statistical Spoken Dialog Systems: A Review. *Proceedings of the IEEE*, 101(5):1160–1179.

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.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of EMNLP*, pages 1393–1398.