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

Nikola Mrkšić¹,², Ivan Vulić¹, Diarmuid Ó Séaghdha², Ira Leviant³
Roi Reichart³, Milica Gašić¹, Anna Korhonen¹, Steve Young¹,²
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 specialised 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-specialised 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 which 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 specialisation.

In this paper we advance the semantic specialisation 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 specialisation by showing the nearest neighbours for three English words across three cross-lingual spaces.
In each case, the vast majority of each words’ neighbours are meaningful synonyms/translations.\(^1\)

While there is a considerable amount of prior research on joint learning of cross-lingual vector spaces (see Sect. 2.2), to the best of our knowledge we are the first to apply semantic specialisation 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 in-

---

\(^1\)Some residual (negative) effects of the distributional hypothesis do persist. For example, *nl_krieken*, which is Dutch for *cherries*, is (presumably) identified as a synonym for *en_morning* due to a song called *a Morning Wish* by Emile Van Krieken.

\(^2\)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.

---

Table 1: Nearest neighbours 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 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 specialised 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-specialised 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 relating to this paper are available at [www.github.com/nmrksic/attract-repel](http://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.
2 Related Work

2.1 Semantic Specialisation

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 specialisation 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 specialisation methods (broadly) fall 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 Databases (PPDB) (Ganitkevitch et al., 2013; Ganitkevitch and Callison-Burch, 2014; Pavlick et al., 2015).

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 ones 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 emphasise 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 Rapoport, 2006) to push away antonymous words in their pattern-based vector space. Ono et al. (2015) combine both approaches, using thesauri and distributional data to train embeddings specialised 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 specialised 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.

---

3The 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 ones 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 specialised 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). For ease of notation, let each word pair $(x_l, x_r)$ in these two sets correspond to a vector pair $(x_l, x_r)$. The optimisation procedure operates over mini-batches $B$, where each of these consists of a set of synonymy pairs $B_S$ (of size $k_1$) and a set of antonymy pairs $B_A$ (of size $k_2$). Let $T_S(B_S) = [(t_1^1, t_1^{l_1}), \ldots, (t_{k_1}^1, t_{k_1}^{l_1})]$ and $T_A(B_A) = [(t_1^2, t_1^{l_2}), \ldots, (t_{k_2}^2, t_{k_2}^{l_2})]$ be the pairs of negative examples for each synonymy and antonymy example pair. These negative examples are chosen from the $2(k_1 + k_2)$ word vectors present in $B_S \cup B_A$:

- For each synonymy pair $(x_l, x_r)$, the negative example pair $(t_r, t_l)$ is chosen from the remaining in-batch vectors so that $t_l$ is the one closest (cosine similarity) to $x_l$ and $t_r$ is closest to $x_r$.

- For each antonymy pair $(x_l, x_r)$, the negative example pair $(t_r, t_l)$ is chosen from the remaining in-batch vectors so that $t_r$ is the one furthest away from $x_l$ and $t_l$ is the one furthest from $x_r$.

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:

$$S(B_S) = \sum_{(x_l, x_r) \in B_S} \left[ \tau \left( \delta_{syn} + x_l t_l - x_l x_r \right) + \tau \left( \delta_{syn} + x_r t_r - x_l x_r \right) \right]$$

where $\tau(x) = \max(0, x)$ is the hinge loss function and $\delta_{syn}$ 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 antony-
mous word pairs away from each other:

\[ A(B_A) = \sum_{(x_i, x_r) \in B_A} \left[ \tau (\delta_{ant} + x_i x_r - x_i t_l) + \tau (\delta_{ant} + x_i x_r - x_i t_r) \right] \]

In addition to these two terms, we include an additional regularisation 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 \( V(B) \) is the set of all word vectors present in the given mini-batch, then:

\[ R(B_S, B_A) = \sum_{x_i \in V(B_S \cup B_A)} \lambda_{\text{reg}} \| \hat{x}_i - x_i \|_2 \]

where \( \lambda_{\text{reg}} \) is the L2 regularisation constant and \( \hat{x}_i \) denotes the original (distributional) word vector for word \( x_i \). The final cost function of the ATTRACT-REPEL algorithm can then be expressed as:

\[ C(B_S, B_A) = S(B_S) + A(B_A) + R(B_S, B_A) \]

**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:

\[ \text{Attract}(S) = \sum_{(x_i, x_r) \in S} \tau (\delta_{\text{syn}} - x_i x_r) \]

where \( S \) is the set of synonymy constraints and \( \delta_{\text{syn}} \) is the (minimum) similarity enforced between synonyms. 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. **Regularisation:**  Counter-fitting preserves distances between pairs of word vectors in the initial vector space, trying to ‘pull’ the words’ neighbourhoods with them as they move to incorporate external knowledge. The radius of this initial neighbourhood introduces an opaque hyperparameter to the procedure. Conversely, ATTRACT-REPEL implements standard L2 regularisation, which ‘pulls’ each vector towards its distributional vector representation.

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

**Optimisation**  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 magnitude of the parameter updates 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 specialised 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 \( \lambda_{\text{reg}} \), the similarity and antonymy margins \( \delta_{\text{sim}} \) and \( \delta_{\text{ant}} \), mini-batch sizes \( k_1 \) and \( k_2 \), and the size of the PPDB constraint set used for each language (larger sizes include more 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.\(^5\)

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

\(^5\)We ran the grid search over \( \lambda_{\text{reg}} \in [10^{-3}, \ldots, 10^{-10}], \delta_{\text{sim}}, \delta_{\text{ant}} \in [0, 0.1, \ldots, 1.0], k_1, k_2 \in [10, 25, 50, 100, 200] \)
We first present our sixteen experimental languages: well-known, high-quality word vector collections: Vietnamese (VI). The first four languages are those (BG), Croatian (HR), Irish (GA), Persian (FA) and (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.

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

![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.](image)

### 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.\(^5\)

### 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 and over the six PPDB sizes for the four SimLex languages. \(\lambda_{reg} = 10^{-6}\), \(\delta_{sim} = 0.6\), \(\delta_{ant} = 0.0\) 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.

**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 inter-language 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.\(^7\)

\(^5\)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.
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. 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. 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 Specialisation

We start from distributional vectors for the SimLex languages: English, German, Italian and Russian. For each language, we first perform semantic specialisation 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 monolingual and cross-lingual specialisation was performed using ATTRACT-REPEL and counter-fitting, in order to conclusively determine which of the two methods exhibited superior performance. Retrofitting and PARA-GRAM 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 similarity (but not antonymy) constraints illustrates the performance range of the two preceding models.

Importance of Initial Vectors

We use three different sets of initial vectors: a) well-known distributional word vector collections (Sect. 4.1); b) distributional vectors trained on the latest Wikipedia dumps; and c) word vectors randomly initialised using the XAVIER initialisation (Glorot and Bengio, 2010).

Specialisation 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. Since each pair contains at least one SimLex language, we can analyse the improvement over monolingual specialisation 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 specialisation 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), re-trained 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 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.

---

8Leviant 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.

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

10Hyperparameters: we used $\delta_{\text{sim}} = 0.6$, $\delta_{\text{ant}} = 0.0$ and $\lambda_{\text{reg}} = 10^{-9}$, 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).
### 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.

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

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

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

### 5.3 Results and Discussion

Table 3 shows the effects of monolingual and cross-lingual semantic specialisation of four well-known distributional vector spaces for the SimLex languages. Monolingual specialisation leads to very strong improvements in the SimLex performance across all languages. Cross-lingual specialisation brings further improvements, with all languages benefiting from sharing the cross-lingual vector space. Italian in particular shows strong evidence of effective transfer, 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 counterfitting: **ATTRACT-REPEL** achieved substantially stronger performance in all experiments. We believe these results conclusively show that the fine-grained updates and L2 regularisation employed by **ATTRACT-REPEL** present a better alternative to the context-insensitive attract/repel terms and pair-wise regularisation employed by counter-fitting.

#### 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 holdout the dataset. This implies that further work which uses these vectors (e.g., (Mrkšić et al., 2016; Recski et al., 2016)) as 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 initialised word vectors; and b) skip-gram with negative sampling vectors trained on the latest Wikipedia dumps. The randomly initialised 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 specialisation: the English performance is surprisingly strong, with other languages suffering more from the lack of initialisation.

---

11The XAVIER initialisation populates the values for each word vector by uniformly sampling from the interval \([-\sqrt{\frac{6}{d}}, +\sqrt{\frac{6}{d}}]\), where \(d\) is the vector dimensionality. This is a typical init method in neural nets research (Goldberg, 2015; Bengio et al., 2013).
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 specialisation. 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.

Table 6: Bilingual semantic specialisation 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.

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 specialisation. However, all vector spaces still exhibit 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 specialisation.

Bilingual Specialisation  Table 5 shows the effect of combining the four original SimLex languages with each other and with twelve other languages (Sect. 4.1). Bilingual specialisation substantially improves over monolingual specialisation 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 ones 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).12

Lower-Resource Languages  The previous experiment indicates that bilingual specialisation 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 specialisation. 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’ language). 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 lead to strongest performance. We believe this shows that bilingual ATTRACT-REPEL specialisation 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

12We 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).
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.

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

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.

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.

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

**Exact Matching as a Bottleneck** Semantic lexicons can be hand-crafted for small dialogue domains. Mrkšić et al. (2016) showed that semantically specialised 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 specialisation 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, 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 recognised purely by their position in the original vector space. Thus, it is essential that deployed word vectors are specialised 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 specialisation 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 specialised vectors; and 3) monolingual subspaces of the cross-lingual semantically specialised EN-DE-IT-RU vectors. For each language, we also compare to the NBT performance achieved using the five state-of-the-art bilingual vector spaces we compared to in Sect. 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

---

13There 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.
vector spaces. The subsequent three rows show the performance of: a) distributional vector spaces; b) their monolingual specialisation; and c) their EN-DE-IT-RU cross-lingual specialisation. 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.

The results in Table 8 show that both types of specialisation 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 specialised 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.

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

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.
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 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 specialises word vectors by jointly injecting mono- and cross-lingual synonymy and antonymy constraints, creating unified cross-lingual vector spaces which achieve 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 low-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.

7.1 Further Work

Our results, especially with DST, emphasise 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


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


