Morph-fitting: Fine-Tuning Word Vector Spaces with Simple Language-Specific Rules

Anonymous ACL submission

Abstract

Morphologically rich languages accentuate two properties of distributional vector space models: 1) the difficulty of inducing accurate representations for low-frequency word forms; and 2) insensitivity to distinct lexical relations that have similar distributional signatures. These effects are detrimental for language understanding systems, which may infer that *inexpensive* is a rephrasing for *expensive* or may not associate *acquire* with *acquires*. In this work, we propose a novel *morph-fitting* procedure which moves past the use of curated semantic lexicons for improving distributional vector spaces. Instead, our method injects morphological constraints generated using simple language-specific rules, pulling *inflectional* forms of the same word close together and pushing *derivational antonyms* far apart. In intrinsic evaluation over four languages, we show that our approach: **1)** improves low-frequency word estimates; and **2)** boosts the semantic quality of the entire word vector collection. Finally, we show that morph-fitted vectors yield large gains in the downstream task of *dialogue state tracking*, highlighting the importance of morphology for tackling long-tail phenomena in language understanding tasks.

1 Introduction

Word representation learning has become a research area of central importance in natural language processing (NLP), with its usefulness demonstrated across many application areas such as parsing (Chen and Manning, 2014), machine translation (Zou et al., 2013), and many others (Turian et al., 2010; Collobert et al., 2011). Most prominent word representation techniques are grounded in the *distributional hypothesis*, relying on word co-occurrence information in large textual corpora (Curran, 2004; Turney and Pantel, 2010; Mikolov et al., 2013; Mnih and Kavukcuoglu, 2013; Levy and Goldberg, 2014; Schwartz et al., 2015, i.a.).

Morphologically rich languages, in which "substantial grammatical information. . . is expressed at word level" (Tsarfaty et al., 2010), pose specific challenges for NLP. This is not always considered when techniques are evaluated on languages such as English or Chinese, which do not have rich morphology. In the case of distributional vector space models, morphological complexity brings two challenges to the fore:

1. Estimating Rare Words: A single lemma can have many different surface realisations. Naively treating each realisation as a separate word leads to sparsity problems and a failure to exploit their shared semantics. On the other hand, lemmatising the entire corpus can obfuscate the differences that exist between different word forms even though they share some aspects of meaning.

2. Embedded Semantics: Morphology can encode semantic relations such as antonymy (e.g. *literate* and *illiterate*, *expensive* and *inexpensive*) or synonymy (*north*, *northern*, *northerly*).

In this work, we tackle the two challenges jointly by introducing a *resource-light* vector space fine-tuning procedure termed *morph-fitting*. The proposed method does not require curated knowledge bases or gold lexicons. Instead, it makes use of the observation that morphology implicitly encodes semantic signals pertaining to synonymy (e.g., German word inflections *katalanisch, katalanischem, katalanischer* denote the same semantic concept in different grammatical roles), and antonymy (e.g., *mature* vs. *immature*), capitalising on the proliferation of word forms in morphologically

en_expensive	**de_teure**	**it_costoso**	**en_slow**	**de_langsam**	**it_lento**	**en_book**	**de_buch**	**it_libro**
costly	teuren	dispendioso	fast	allmählich	lentissimo	books	sachbuch	romanzo
costlier	kostspielige	remunerativo	slower	rasch	lenta	memoir	buches	racconto
cheaper	aufwändige	redditizio	slower	gemächlich	inesorabile	novel	romandebüt	volumetto
prohibitively	kostenintensive	rischioso	slowed	schnell	rapidissimo	storybooks	büchlein	saggio
pricey	aufwendige	costosa	slowing	explosionsartig	graduale	blurb	pamphlet	ecclesiaste
expensiveness	teures	costosa	slow	langsamer	lenti	booked	bücher	libri
costly	teuren	costose	slowing	langsames	lente	rebook	büch	libra
costlier	teurem	costosi	slowed	langsame	lenta	booking	büche	librare
ruinously	teurer	dispendioso	slowness	langsamem	veloce	rebooked	büches	libre
unaffordable	teurerer	dispendiose	slows	langsamen	rapido	books	büchen	librano

Table 1: The nearest neighbours of three example words (*expensive*, *slow* and *book*) in English, German and Italian before (top) and after (bottom) morph-fitting.

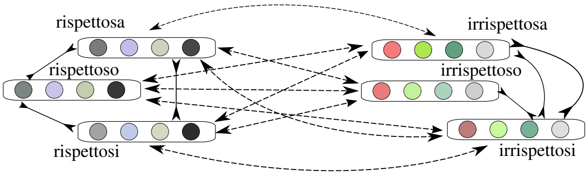


Figure 1: *Morph-fitting* in Italian. Representations for *rispettoso*, *rispettosa*, *rispettosi* (EN: *respectful*), are pulled closer together in the vector space (solid lines; ATTRACT constraints). At the same time, the model pushes them away from their antonyms (dashed lines; REPEL constraints) *irrispettoso*, *irrispettosa*, *irrispettosi* (EN: *disrespectful*), obtained through morphological affix transformation.

rich languages. Formalised as an instance of the post-processing *semantic specialisation* paradigm (Faruqui et al., 2015; Mrkšić et al., 2016), morph-fitting is steered by a set of linguistic constraints derived from simple language-specific rules which describe (a subset of) morphological processes in a language. The constraints emphasise similarity on one side (e.g., by extracting *morphological* synonyms), and antonymy on the other (by extracting *morphological* antonyms), see Fig. 1 and Tab. 2.

The key idea of the fine-tuning process is to pull synonymous examples described by the constraints closer together in a transformed vector space, while at the same time pushing antonymous examples away from each other. The explicit post-hoc injection of morphological constraints enables: **a)** estimating more accurate vectors for low-frequency words if they are described by the constraints containing their relation with high-frequency words,[1] thus tackling the data sparsity problem; and **b)** specialising the distributional space to distinguish between similarity and association, thus supporting language understanding applications such as *dialogue state tracking* (DST).

As a post-processor, morph-fitting allows the integration of morphological rules with any distributional vector space in any language: it treats an input distributional word vector space as a black box and fine-tunes it so that the transformed space reflects the knowledge coded in the input morphological constraints (e.g., Italian words *rispettoso* and *irrispetosa* should be far apart in the transformed vector space, see Fig. 1). Tab. 1 illustrates the effects of morph-fitting by qualitative examples in three languages: the vast majority of nearest neighbours are "morphological" synonyms.

We demonstrate the efficacy of morph-fitting in four languages (English, German, Italian, Russian), yielding large and consistent improvements on benchmarking word similarity evaluation sets such as SimLex-999 (Hill et al., 2015), its multilingual extension (Leviant and Reichart, 2015), and SimVerb-3500 (Gerz et al., 2016). The improvements are reported for all four languages, and with a variety of input distributional spaces, verifying the robustness of the approach.

We then show that incorporating morph-fitted vectors into a state-of-the-art neural-network DST model results in improved tracking performance, especially for morphologically rich languages. We report an improvement of 4% on Italian, and 6% on German when using *morph-fitted* vectors instead of the distributional ones, setting a new state-of-the-art DST performance for the two datasets.[2]

2 Morph-fitting: Methodology

Preliminaries In this work, we focus on four languages with varying levels of morphological complexity: English (EN), German (DE), Italian (IT), and Russian (RU). These correspond to languages in the Multilingual SimLex-999 dataset. Vocabularies W_{en}, W_{de}, W_{it}, W_{ru} are compiled by retaining all word forms from the four Wikipedias with

[1] For instance, the vector for the word *katalanischem* which occurs only 9 times in the German Wikipedia will be pulled closer to the more reliable vectors for *katalanisch* and *katalanischer*, with frequencies of 2097 and 1383 respectively.

[2] There are no readily available DST datasets for Russian.

word frequency over 10, see Tab. 3. We then query these (large) vocabularies using a set of simple language-specific *if-then-else* rules to extract sets of linguistic constraints, see Tab. 2.[3] These constraints (Sect. 2.2) are used as input for the vector space post-processing ATTRACT-REPEL algorithm (outlined in Sect. 2.1).

2.1 The ATTRACT-REPEL Model

The ATTRACT-REPEL model is an extension of PARAGRAM, proposed by Wieting et al. (2015). It provides a generic framework for incorporating *similarity* (e.g. *successful* and *accomplished*) and *antonymy* constraints (e.g. *nimble* and *clumsy*) into pre-trained word vectors. Given the initial vector space and collections of ATTRACT and REPEL constraints A and R, the model gradually modifies the space to bring the designated word vectors closer together or further apart. The method's cost function consists of three terms. The first term pulls the ATTRACT examples $(x_l, x_r) \in A$ closer together. If B_A denotes the current mini-batch of ATTRACT examples, this term can be expressed as:

$$\begin{aligned} A(\mathcal{B}_A) = \sum_{(x_l,x_r)\in\mathcal{B}_A} & (ReLU\,(\delta_{att} + \mathbf{x}_l\mathbf{t}_l - \mathbf{x}_l\mathbf{x}_r) \\ + & \; ReLU\,(\delta_{att} + \mathbf{x}_r\mathbf{t}_r - \mathbf{x}_l\mathbf{x}_r)) \end{aligned}$$

where δ_{att} is the similarity margin which determines how much closer synonymous vectors should be to each other than to each of their respective negative examples. $ReLU(x) = \max(0, x)$ is the standard rectified linear unit (Nair and Hinton, 2010). The 'negative' example $\mathbf{t}_i$ for each word x_i in any ATTRACT pair is the word vector *closest* to $\mathbf{x}_i$ among the examples in the current mini-batch (distinct from its target synonym and $\mathbf{x}_i$ itself). This means that this term forces synonymous words from the in-batch ATTRACT constraints to be closer to one another than to any other word in the current mini-batch.

The second term pushes antonyms away from each other. If $(x_l, x_r) \in B_R$ is the current mini-batch of REPEL constraints, this term is:

$$\begin{aligned} R(\mathcal{B}_R) = \sum_{(x_l,x_r)\in\mathcal{B}_R} & (ReLU\,(\delta_{rpl} + \mathbf{x}_l\mathbf{x}_r - \mathbf{x}_l\mathbf{t}_r) \\ + & \; ReLU\,(\delta_{rpl} + \mathbf{x}_l\mathbf{x}_r - \mathbf{x}_r\mathbf{t}_r)) \end{aligned}$$

English	German	Italian
(discuss, discussed)	(schottisch, schottischem)	(golfo, golfi)
(laugh, laughing)	(damalige, damaligen)	(minato, minata)
(pacifist, pacifists)	(kombiniere, kombinierte)	(mettere, metto)
(evacuate, evacuated)	(schweigt, schweigst)	(crescono, cresci)
(evaluate, evaluates)	(hacken, gehackt)	(crediti, credite)
(dressed, undressed)	(stabil, unstabil)	(abitata, inabitato)
(similar, dissimilar)	(geformtes, ungeformt)	(realtà, irrealtà)
(formality, informality)	(relevant, irrelevant)	(attuato, inattuato)

Table 2: Example synonymous (inflectional; top) and antonymous (derivational; bottom) constraints.

	$\lvert W \rvert$	$\lvert A \rvert$	$\lvert R \rvert$
English	1,368,891	231,448	45,964
German	1,216,161	648,344	54,644
Italian	541,779	278,974	21,400
Russian	950,783	408,400	32,174

Table 3: Vocabulary sizes and counts of ATTRACT (A) and REPEL (R) constraints.

In this case, each word's 'negative' example is the (in-batch) word vector furthest away from it (and distinct from the word's target antonym). The intuition is that we want antonymous words from the input REPEL constraints to be *further away* from each other than from any other word in the current mini-batch; δ_{rpl} is now the *repel* margin.

The final term of the cost function serves to retain the abundance of semantic information encoded in the starting distributional space. If $\mathbf{x}_i^{init}$ is the initial distributional vector and $V(\mathcal{B})$ is the set of all vectors present in the given mini-batch, this term (per mini-batch) is expressed as:

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

where λ_{reg} is the L2 regularisation constant.[4] This term effectively *pulls* word vectors towards their initial (distributional) values, ensuring that relations encoded in initial vectors persist as long as they do not contradict the newly injected ones.

2.2 Language-Specific Rules and Constraints

Semantic Specialisation with Constraints The fine-tuning ATTRACT-REPEL procedure is entirely driven by the input ATTRACT and REPEL sets of constraints. These can be extracted from a variety of semantic databases such as WordNet (Fellbaum, 1998), the Paraphrase Database (Ganitkevitch et al., 2013; Pavlick et al., 2015), or BabelNet (Navigli and Ponzetto, 2012; Ehrmann et al., 2014) as done in prior work (Faruqui et al., 2015; Wieting et al.,

[3] A native speaker is able to easily come up with these sets of morphological rules (or at least with a reasonable subset of rules) without any linguistic training. What is more, the rules for DE, IT, and RU were created by non-native, non-fluent speakers with a limited knowledge of the three languages, exemplifying the simplicity and portability of the approach.

[4] We use hyperparameter values $\delta_{att} = 0.6$, $\delta_{rpl} = 0.0$, $\lambda_{reg} = 10^{-9}$ from prior work without fine-tuning. We train all models for 10 epochs with AdaGrad (Duchi et al., 2011).

2015; Mrkšić et al., 2016, i.a.). In this work, we investigate another option: extracting constraints *without* curated knowledge bases in a spectrum of languages by exploiting inherent language-specific properties related to linguistic morphology. This relaxation ensures a wider portability of ATTRACT-REPEL to languages and domains without readily available or adequate resources.

Extracting ATTRACT Pairs For the ATTRACT constraints, we focus on *inflectional* rather than on *derivational morphology* rules as the former preserve the full meaning of a word, modifying it only to reflect grammatical roles (e.g., verb tense, case markers; *(read, reads)*).[5] This choice is guided by our intent to fine-tune the original vector space to improve the embedded semantic relations.

We define two rules for English, widely recognised as morphologically simple (Avramidis and Koehn, 2008; Cotterell et al., 2016). These are: **(R1)** *if* $w_1, w_2 \in W_{en}$, where $w_2 = w_1$ + *ing/ed/s*, *then* add (w_1, w_2) and (w_2, w_1) to the set of ATTRACT constraints A. This rule yields pairs such as *(look, looks), (look, looking), (look, looked)*.

If $w[: -1]$ is a function which strips the last character from word w, the second rule is: **(R2)** *if* w_1 ends with the letter *e and* $w_1 \in W_{en}$ *and* $w_2 \in W_{en}$, where $w_2 = w_1[: -1]$ + *ing/ed/s*, *then* add (w_1, w_2) and (w_2, w_1) to A. This creates pairs such as *(create, creates), (create, creating)* and *(create, created)*. Naturally, introducing more sophisticated rules is possible in order to cover for other special cases and morphological irregularities (e.g., *sweep / swept*), but in all our EN experiments, A is based on the two simple EN rules R1 and R2.

The other three languages, with more complicated morphology, yield a larger number of rules. In Italian, we rely on the sets of rules spanning: (1) regular formation of plural (*libro / libri*); (2) regular verb conjugation (*aspettare / aspettiamo*); (3) regular formation of past participle (*aspettare / aspettato*); and (4) rules regarding grammatical gender (*bianco / bianca*). Besides these, another set of rules is used for German and Russian: (5) regular declension (e.g., *asiatisch / asiatischem*).

Extracting REPEL Pairs As another source of implicit semantic signals, W also contains words which represent *derivational antonyms*: e.g., two words that denote concepts with opposite meanings, generated through a derivational process. We use a standard set of EN "antonymy" prefixes: AP_{en} = *{dis, il, un, in, im, ir, mis, non, anti}* (Fromkin et al., 2013). *If* $w_1, w_2 \in W_{en}$, where w_2 is generated by adding a prefix from AP_{en} to w_1, *then* (w_1, w_2) and (w_2, w_1) are added to the set of REPEL constraints R. This rule generates pairs such as *(advantage, disadvantage)* and *(regular, irregular)*. An additional rule replaces the suffix *-ful* with *-less*, extracting antonyms such as *(careful, careless)*.

Following the same principle, we use AP_{de} = *{un, nicht, anti, ir, in, miss}*, AP_{it} = *{in, ir, im, anti}*, and AP_{ru} = *{*не, анти*}*. For instance, this generates an IT pair *(rispettoso, irrispettoso)* (see Fig. 1). For DE, we use another rule targeting suffix replacement: *-voll* is replaced by *-los*.

We further expand the set of REPEL constraints by transitively combining antonymy pairs from the previous step with inflectional ATTRACT pairs. This step yields additional constraints such as *(rispettosa, irrispettosi)* (see Fig. 1). The final A and R constraint counts are given in Tab. 3. The full sets of rules are available as supplemental material.

3 Experimental Setup

Training Data and Setup For each of the four languages we train the skip-gram with negative sampling (SGNS) model (Mikolov et al., 2013) on the latest Wikipedia dump of each language. We induce 300-dimensional word vectors, with the frequency cut-off set to 10. The vocabulary sizes $|W|$ for each language are provided in Tab. 3.[6] We label these collections of vectors SGNS-LARGE.

Other Starting Distributional Vectors We also analyse the impact of *morph-fitting* on other collections of well-known EN word vectors. These vectors have varying vocabulary coverage and are trained with different architectures. We test standard distributional models: Common-Crawl GloVe (Pennington et al., 2014), SGNS vectors (Mikolov et al., 2013) with various contexts (*BOW* = bag-of-words; *DEPS* = dependency contexts), and training data (*PW* = Polyglot Wikipedia from Al-Rfou

[5] The core difference between *inflectional* and *derivational morphology* may be summarised in a few lines as follows: the former refers to a set of processes through which the word form expresses meaningful syntactic information, e.g., verb tense, without any change to the semantics of the word. On the other hand, the latter refers to the formation of new words with semantic shifts in meaning (Schone and Jurafsky, 2001; Haspelmath and Sims, 2013; Lazaridou et al., 2013; Zeller et al., 2013; Cotterell and Schütze, 2017).

[6] Other SGNS parameters were set to standard values (Baroni et al., 2014; Vulić and Korhonen, 2016b): 15 epochs, 15 negative samples, global learning rate: .025, subsampling rate: $1e-4$. Similar trends in results persist with $d = 100, 500$.

et al. (2013); *8B* = 8 billion token `word2vec` corpus), following (Levy and Goldberg, 2014) and (Schwartz et al., 2015). We also test the symmetric-pattern based vectors of Schwartz et al. (2016) (*SymPat-Emb*), count-based PMI-weighted vectors reduced by SVD (Baroni et al., 2014) (*Count-SVD*), a model which replaces the context modelling function from CBOW with bidirectional LSTMs (Melamud et al., 2016) (*Context2Vec*), and two sets of EN vectors trained by injecting multilingual information: *BiSkip* (Luong et al., 2015) and *MultiCCA* (Faruqui and Dyer, 2014). We also experiment with a selection of standard distributional spaces in other languages from prior work (Dinu et al., 2015; Luong et al., 2015; Vulić and Korhonen, 2016a).

Morph-fixed Vectors A baseline which utilises an equal amount of knowledge as morph-fitting, termed *morph-fixing*, fixes the vector of each word to the distributional vector of its most frequent inflectional synonym, tying the vectors of low-frequency words to their more frequent inflections. For each word w_1, we construct a set of $M+1$ words $W_{w_1} = \{w_1, w'_1, \ldots, w'_M\}$ consisting of the word w_1 itself and all M words which co-occur with w_1 in the ATTRACT constraints. We then choose the word w'_{max} from the set W_{w_1} with the maximum frequency in the training data, and fix all other word vectors in W_{w_1} to its word vector. The morph-fixed vectors (MFIX) serve as our primary baseline, as they outperformed another straightforward baseline based on *stemming* across all of our intrinsic and extrinsic experiments.

Morph-fitting Variants We analyse two variants of morph-fitting: (1) using ATTRACT constraints only (MFIT-A), and (2) using both ATTRACT and REPEL constraints (MFIT-AR).[7]

4 Intrinsic Evaluation: Word Similarity

Evaluation Setup and Datasets The first set of experiments intrinsically evaluates *morph-fitted* vector spaces on word similarity benchmarks, using Spearman's rank correlation as the evaluation metric. First, we use the SimLex-999 dataset, as well as SimVerb-3500, a recent EN verb pair similarity dataset providing similarity ratings for 3,500 verb pairs.[8] SimLex-999 was translated to DE, IT, and RU by Leviant and Reichart (2015), and they crowd-sourced similarity scores from native speakers. We use this dataset for our multilingual evaluation.[9]

Morph-SimLex We also introduce a synthetic dataset based on multilingual SimLex, termed *Morph-SimLex*. Since the original sets contain only word lemmas, they are unable to evaluate whether a representation model improves vectors for all synonymous word inflections. Therefore, we enrich the sets of pairs using the same set of ATTRACT rules from Sect. 2.2. In short, given a word pair (w_1, w_2) with a SimLex score $sl_{1,2}$, we again construct sets $W_{w_1} = \{w_1, w'_1, \ldots, w'_M\}$ and $W_{w_2} = \{w_2, w''_1, \ldots, w''_N\}$, where W_{w_1} consists of w_1 and all words which co-occur with w_1 in the A constraints; the same holds for W_{w_2}. Morph-SimLex pairs are then generated by taking the Cartesian product between W_{w_1} and W_{w_2}, and assigning the same score $sl_{1,2}$ to each such pair. The final dataset is constructed by repeating the procedure for each of the 999 SimLex pairs, yielding 13,213 EN pairs, 17,021 DE pairs, 18,281 IT pairs, and 10,289 RU pairs. We make this dataset available in hope it can aid further research on improving morphological relations in vector spaces.

Morph-fitting EN Word Vectors As the first experiment, we morph-fit a wide spectrum of EN distributional vectors induced by various architectures (see Sect. 3). The results on SimLex and SimVerb are summarised in Tab. 4. The results with EN SGNS-LARGE vectors are shown in Fig. 2a. Morph-fitted vectors bring consistent improvement across all experiments, regardless of the quality of the initial distributional space. This finding confirms that the method is robust: its effectiveness does not depend on the architecture used to construct the initial space. To illustrate the improvements, note that the best score on SimVerb for a model trained on running text is achieved by *Context2vec* ($\rho = 0.388$); injecting morphological constraints into this vector space results in a gain of 7.1 ρ points.

Experiments on Other Languages We next extend our experiments to other languages, testing both morph-fitting variants. The results are sum-

[7] We also tried using another post-processing model (Mrkšić et al., 2016) in lieu of ATTRACT-REPEL. However, this model was computationally intractable with SGNS-LARGE vectors. Moreover, it was consistently outperformed by ATTRACT-REPEL on vector spaces with smaller vocabularies.

[8] Unlike other gold standard resources such as WordSim-353 (Finkelstein et al., 2002) or MEN (Bruni et al., 2014), SimLex and SimVerb provided explicit guidelines to discern between semantic similarity and association, so that related but non-similar words (e.g. *cup* and *coffee*) have a low rating.

[9] Since Leviant and Reichart (2015) re-scored the original EN SimLex, we use their EN SimLex version for consistency.

	Evaluation	
Vectors	SimLex-999	SimVerb-3500
1. SG-BOW2-PW (300) (Mikolov et al., 2013)	.339 → **.439**	.277 → **.381**
2. GloVe-6B (300) (Pennington et al., 2014)	.324 → **.438**	.286 → **.405**
3. Count-SVD (500) (Baroni et al., 2014)	.267 → **.360**	.199 → **.301**
4. SG-DEPS-PW (300) (Levy and Goldberg, 2014)	.376 → **.434**	.313 → **.418**
5. SG-DEPS-8B (500) (Bansal et al., 2014)	.373 → **.441**	.356 → **.473**
6. MultiCCA-EN (512) (Faruqui and Dyer, 2014)	.314 → **.391**	.296 → **.354**
7. BiSkip-EN (256) (Luong et al., 2015)	.276 → **.356**	.260 → **.333**
8. SG-BOW2-8B (500) (Schwartz et al., 2015)	.373 → **.440**	.348 → **.441**
9. SymPat-Emb (500) (Schwartz et al., 2016)	.381 → **.442**	.284 → **.373**
10. Context2Vec (600) (Melamud et al., 2016)	.371 → **.440**	.388 → **.459**

Table 4: The impact of morph-fitting (MFIT-AR used) on a representative set of EN vector space models. All results show the Spearman's ρ correlation before and after morph-fitting. The numbers in parentheses refer to the vector dimensionality.

Vectors	Distrib.	MFIT-A	MFIT-AR
EN: GloVe-6B (300)	.324	.376	**.438**
EN: SG-BOW2-PW (300)	.339	.385	**.439**
DE: SG-DEPS-PW (300) (Vulić and Korhonen, 2016a)	.267	.318	**.325**
DE: BiSkip-DE (256) (Luong et al., 2015)	.354	.414	**.421**
IT: SG-DEPS-PW (300) (Vulić and Korhonen, 2016a)	.237	.351	**.391**
IT: CBOW5-Wacky (300) (Dinu et al., 2015)	.363	.417	**.446**

Table 5: Results on multilingual SimLex-999 (EN, DE, and IT) with two morph-fitting variants.

marised in Tab. 5, while Fig. 2a-2d show results for the morph-fitted SGNS-LARGE vectors. These scores confirm the effectiveness and robustness of morph-fitting across languages, suggesting that the idea of fitting to morphological constraints is indeed language-agnostic, given the set of language-specific rule-based constraints. Fig. 2 also demonstrates that the morph-fitted vector spaces consistently outperform the morph-fixed ones.

Morph-SimLex performance across all languages shows even stronger relative gains over distributional and morph-fixed vectors. The original SimLex dataset only contains word lemmas. Consequently, it fails to penalise word vector collections with bad estimates of less-frequent word forms. The comparison between MFIT-A and MFIT-AR indicates that both sets of constraints are important for the fine-tuning process: while MFIT-A already yields consistent gains over the initial spaces, a further refinement can be achieved by also incorporating the antonymous REPEL constraints.

5 Downstream Task: Dialogue State Tracking (DST)

Goal-oriented dialogue systems provide conversational interfaces for tasks such as booking flights or finding restaurants. In *slot-based* systems, application domains are specified using *ontologies* that define the search constraints which users can express. An ontology consists of a number of *slots* and their assorted *slot values*. In a *restaurant search* domain, sets of slot-values could include PRICE = [*cheap, expensive*] or FOOD = [*Thai, Indian, ...*]. The DST model is the first component of modern dialogue pipelines (Young, 2010). It serves to capture the intents expressed by the user at each dialogue turn and update the *belief state*. This is the system's internal estimate of the user's goals, used by the downstream *dialogue manager* to choose the system response. The following example shows the true dialogue state in a multi-turn dialogue:

User: What's good in the southern part of town?
`inform(area=south)`

System: Vedanta is the top-rated Indian place.

User: How about something cheaper?
`inform(area=south, price=cheap)`

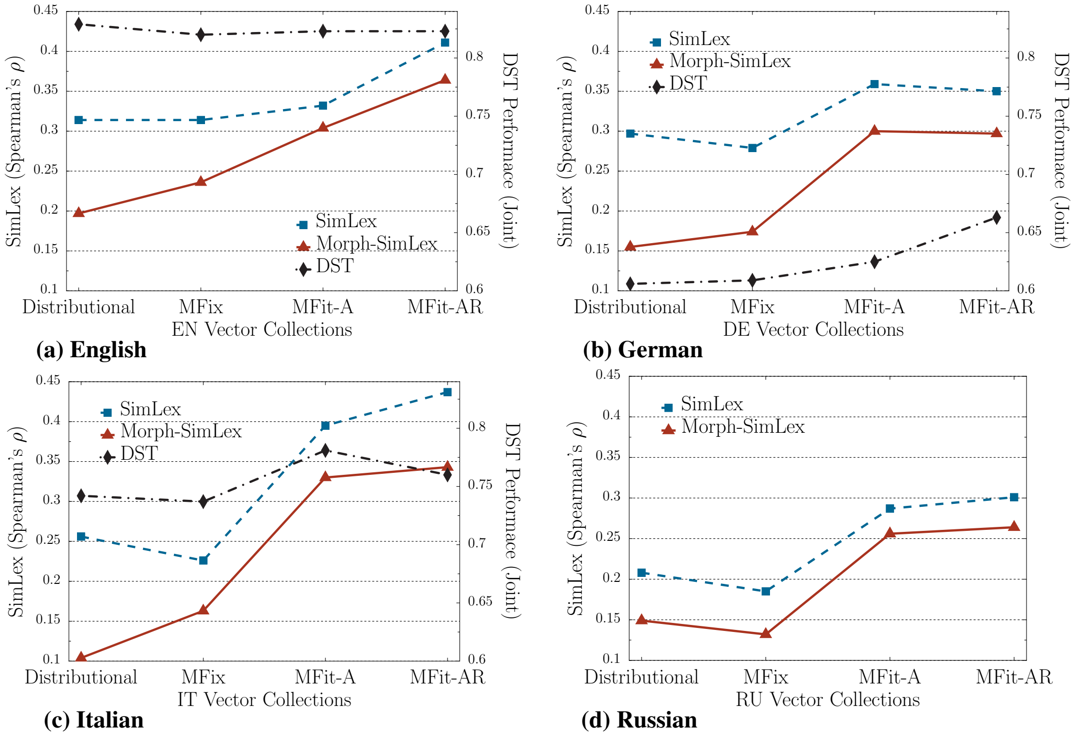
System: Seven Days is very popular. Great hot pot.

User: What's the address?
`inform(area=south, price=cheap);`
`request(address)`

System: Seven Days is at 66 Regent Street.

The Dialogue State Tracking Challenge (DSTC) shared task series formalised the evaluation and provided labelled DST datasets (Henderson et al., 2014a,b; Williams et al., 2016). While a plethora of DST models are available based on, e.g., handcrafted rules (Wang et al., 2014) or conditional random fields (Lee and Eskenazi, 2013), the recent DST methodology has seen a shift towards neural-network architectures (Henderson et al., 2014c; Mrkšić et al., 2015; Liu and Perez, 2017, i.a.)

Model: Neural Belief Tracker To detect intents in user utterances, most existing models rely on either (or both): **1)** Spoken Language Understanding models which require large amounts of annotated training data; or **2)** hand-crafted, domain-specific lexicons which try to capture lexical and morphological variation. The Neural Belief Tracker (NBT) is a novel DST model which overcomes both issues



(a) English

(b) German

(c) Italian

(d) Russian

Figure 2: An overview of the results (Spearman's ρ correlation) for four languages on SimLex-999 (blue squares), Morph-SimLex-999 (red triangles), and the downstream DST performance (black diamonds) using SGNS-LARGE vectors ($d = 300$), see Tab. 3 and Sect. 3. The left y axis measures the intrinsic word similarity performance, while the right y axis provides the scale for the DST performance.

by reasoning purely over pre-trained word vectors (Mrkšić et al., 2016). The NBT learns to compose these vectors into intermediate utterance and context representations. These are then used to decide which of the ontology-defined intents (goals) have been expressed by the user. The NBT model keeps word vectors *fixed* during training, so that unseen, yet related words can be mapped to the right intent at test time (e.g. *northern* to *north*).

Data: Multilingual WOZ 2.0 Dataset Our DST evaluation is based on the WOZ dataset, released by Wen et al. (2017). In this Wizard-of-Oz setup, two Amazon Mechanical Turk workers assumed the role of the user and the system asking/providing restaurant information. Users typed instead of speaking, removing the need to deal with noisy speech recognition. In DSTC datasets, users would quickly adapt to the system's inability to deal with complex queries. Conversely, the WOZ setup allowed them to use sophisticated language. The WOZ 2.0 release expanded the dataset to 1,200 dialogues (Mrkšić et al., 2016). In this work, we use translations of this dataset to Italian and German, provided by the authors of the original dataset.

Evaluation Setup The principal metric we use to measure DST performance is *joint goal accuracy*, which represents the proportion of test set dialogue turns where all user goals expressed up to that point of the dialogue were decoded correctly (Henderson et al., 2014a). The NBT models for EN, DE and IT are trained using four variants of the SGNS-LARGE vectors: **1)** the initial distributional vectors; **2)** *morph-fixed*; **3)** and **4)** the two variants of *morph-fitted* vectors (see Sect. 3).

Results and Discussion The diamond-dashed lines (against the right axes) in Fig. 2 show the DST performance of NBT models making use of the four vector collections. IT and DE benefit from both kinds of *morph-fitting*: IT performance increases 74.1 → 78.1 (MFIT-A) and DE performance rises even more: 60.6 → 66.3 (MFIT-AR), setting a new state-of-the-art score for both languages. The *morph-fixed* vectors do not enhance DST performance, probably because fixing word vectors to their highest frequency inflectional form eliminates useful semantic content encoded in the original vectors. On the other hand, morph-fitting makes use of this information, supplementing it with semantic relations between different morphological forms. These conclusions are in line with the SimLex and Morph-SimLex gains, where morph-fitting outperforms distributional and *morph-fixed* vectors.

English performance shows little variation across the four word vector collections investigated here. This corroborates our intuition that, as a morphologically simpler language, English stands to gain less from fine-tuning the morphological variation for downstream applications. This result again points at the discrepancy between intrinsic and extrinsic evaluation: the considerable gains in SimLex performance do not necessarily induce similar gains in downstream performance.

6 Related Work

Semantic Specialisation A standard approach to incorporating external information into vector spaces is to pull the representations of similar words closer together. Some models integrate such constraints into the training procedure, modifying the prior or the regularisation (Yu and Dredze, 2014; Xu et al., 2014; Bian et al., 2014; Kiela et al., 2015), or using a variant of the SGNS-style objective (Liu et al., 2015; Osborne et al., 2016). Another class of models, popularly termed *retrofitting*, injects lexical knowledge from available semantic databases (e.g., WordNet, PPDB) into pre-trained word vectors (Faruqui et al., 2015; Jauhar et al., 2015; Wieting et al., 2015; Nguyen et al., 2016; Mrkšić et al., 2016). Morph-fitting falls into the latter category. However, instead of resorting to curated knowledge bases, and experimenting solely with English, we show that the *morphological richness* of any language can be exploited as a source of inexpensive supervision for fine-tuning vector spaces, at the same time specialising them to better reflect true semantic similarity.

Word Vectors and Morphology The use of morphological resources to improve the representations of morphemes and words is an active area of research. The majority of proposed architectures encode morphological information, provided either as gold standard morphological resources (Sylak-Glassman et al., 2015) such as CELEX (Baayen et al., 1995) or as an external analyser such as Morfessor (Creutz and Lagus, 2007), along with distributional information jointly at *training* time in the language modelling (LM) objective (Luong et al., 2013; Botha and Blunsom, 2014; Qiu et al., 2014; Cotterell and Schütze, 2015; Bhatia et al., 2016, i.a.). The key idea is to learn a morphological composition function (Lazaridou et al., 2013; Cotterell and Schütze, 2017) which synthesises the representation of a word given the representations of its constituent morphemes. Contrary to our work, these models typically coalesce all lexical relations.

Another class of models, operating at the character level, shares a similar methodology: such models compose token-level representations from subcomponent embeddings (subwords, morphemes, or characters) (dos Santos and Zadrozny, 2014; Ling et al., 2015; Cao and Rei, 2016; Kim et al., 2016; Wieting et al., 2016; Verwimp et al., 2017, i.a.).

In contrast to prior work, our model *decouples* the use of morphological information, now provided in the form of inflectional and derivational rules transformed into linguistic constraints, from the actual training. This pipelined approach results in a simpler, more portable model. In spirit, our work is similar to Cotterell et al. (2016), who formulate the idea of post-training specialisation in a generative Bayesian framework. Their work uses gold morphological lexicons; we show that competitive performance can be achieved using a non-exhaustive set of simple rules. Our framework facilitates the inclusion of *antonyms* at no extra cost and naturally extends to constraints from other sources (e.g., WordNet) in future work. Another practical difference is that we focus on similarity and evaluate morph-fitting in a well-defined downstream task where the artefacts of the distributional hypothesis are known to prompt statistical system failures.

7 Conclusion and Future Work

We have presented a novel *morph-fitting* method which injects morphological knowledge in the form of linguistic constraints into word vector spaces. The method makes use of implicit semantic signals encoded in inflectional and derivational rules which describe the morphological processes in a language. The results in intrinsic word similarity tasks show that *morph-fitting* improves vector spaces induced by distributional models across four languages. Finally, we have shown that the use of *morph-fitted* vectors boosts the performance of downstream language understanding models which rely on word representations as features, especially for morphologically rich languages such as German.

Future work will focus on other potential sources of morphological knowledge (Soricut and Och, 2015), porting the framework to other morphologically rich languages and downstream tasks, and on further refinements of the post-processing algorithm and constraints selection.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of CoNLL*. pages 183–192. http://www.aclweb.org/anthology/W13-3520.

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of ACL*. pages 763–770. http://www.aclweb.org/anthology/P/P08/P08-1087.

Harald R. Baayen, Richard Piepenbrock, and Hedderik van Rijn. 1995. The CELEX lexical data base on CD-ROM .

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of ACL*. pages 809–815. http://www.aclweb.org/anthology/P14-2131.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*. pages 238–247. http://www.aclweb.org/anthology/P14-1023.

Parminder Bhatia, Robert Guthrie, and Jacob Eisenstein. 2016. Morphological priors for probabilistic neural word embeddings. In *Proceedings of EMNLP*. pages 490–500. https://aclweb.org/anthology/D16-1047.

Jiang Bian, Bin Gao, and Tie-Yan Liu. 2014. Knowledge-powered deep learning for word embedding. In *Proceedings of ECML-PKDD*. pages 132–148. https://doi.org/10.1007/978-3-662-44848-9_9.

Jan A. Botha and Phil Blunsom. 2014. Compositional morphology for word representations and language modelling. In *Proceedings of ICML*. pages 1899–1907. http://jmlr.org/proceedings/papers/v32/botha14.html.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research* 49:1–47. https://doi.org/10.1613/jair.4135.

Kris Cao and Marek Rei. 2016. A joint model for word embedding and word morphology. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. pages 18–26. http://aclweb.org/anthology/W/W16/W16-1603.

Danqi Chen and Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP*. pages 740–750. http://www.aclweb.org/anthology/D14-1082.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12:2493–2537. http://dl.acm.org/citation.cfm?id=1953048.2078186.

Ryan Cotterell and Hinrich Schütze. 2015. Morphological word-embeddings. In *Proceedings of NAACL-HLT*. pages 1287–1292. http://www.aclweb.org/anthology/N15-1140.

Ryan Cotterell and Hinrich Schütze. 2017. Joint semantic synthesis and morphological analysis of the derived word. *Transactions of the ACL (to appear)* https://arxiv.org/abs/1701.00946.

Ryan Cotterell, Hinrich Schütze, and Jason Eisner. 2016. Morphological smoothing and extrapolation of word embeddings. In *Proceedings of ACL*. pages 1651–1660. http://www.aclweb.org/anthology/P16-1156.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *TSLP* 4(1):3:1–3:34. http://doi.acm.org/10.1145/1217098.1217101.

James Curran. 2004. *From Distributional to Semantic Similarity*. Ph.D. thesis, School of Informatics, University of Edinburgh. http://hdl.handle.net/1842/563.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of ICLR (Workshop Papers)*. http://arxiv.org/abs/1412.6568.

Cícero Nogueira dos Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *Proceedings of ICML*. pages 1818–1826. http://jmlr.org/proceedings/papers/v32/santos14.html.

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. http://dl.acm.org/citation.cfm?id=2021068.

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. http://www.lrec-conf.org/proceedings/lrec2014/summaries/810.html.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL-HLT*. pages 1606–1615. http://www.aclweb.org/anthology/N15-1184.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of EACL*. pages 462–471. http://www.aclweb.org/anthology/E14-1049.

Christiane Fellbaum. 1998. *WordNet*. https://mitpress.mit.edu/books/wordnet.

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. https://doi.org/10.1145/503104.503110.

Victoria Fromkin, Robert Rodman, and Nina Hyams. 2013. *An Introduction to Language, 10th Edition*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of NAACL-HLT*. pages 758–764. http://www.aclweb.org/anthology/N13-1092.

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. https://aclweb.org/anthology/D16-1235.

Martin Haspelmath and Andrea Sims. 2013. *Understanding morphology*.

Matthew Henderson, Blaise Thomson, and Jason D. Wiliams. 2014a. The Second Dialog State Tracking Challenge. In *Proceedings of SIGDIAL*. pages 263–272. http://aclweb.org/anthology/W/W14/W14-4337.pdf.

Matthew Henderson, Blaise Thomson, and Jason D. Wiliams. 2014b. The Third Dialog State Tracking Challenge. In *Proceedings of IEEE SLT*. pages 324–329. https://doi.org/10.1109/SLT.2014.7078595.

Matthew Henderson, Blaise Thomson, and Steve Young. 2014c. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of SIGDIAL*. pages 292–299. http://aclweb.org/anthology/W/W14/W14-4340.pdf.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* 41(4):665–695. https://doi.org/10.1162/COLI_a_00237.

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. http://www.aclweb.org/anthology/N15-1070.

Douwe Kiela, Felix Hill, and Stephen Clark. 2015. Specializing word embeddings for similarity or relatedness. In *Proceedings of EMNLP*. pages 2044–2048. http://aclweb.org/anthology/D15-1242.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of AAAI*. pages 2741–2749.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of ACL*. pages 1517–1526. http://www.aclweb.org/anthology/P13-1149.

Sungjin Lee and Maxine Eskenazi. 2013. Recipe for building robust spoken dialog state trackers: Dialog State Tracking Challenge system description. In *Proceedings of SIGDIAL*. pages 414–422. http://aclweb.org/anthology/W/W13/W13-4066.pdf.

Ira Leviant and Roi Reichart. 2015. Separated by an un-common language: Towards judgment language informed vector space modeling. *CoRR* abs/1508.00106. http://arxiv.org/abs/1508.00106.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL*. pages 302–308. http://www.aclweb.org/anthology/P14-2050.

Wang Ling, Chris Dyer, Alan W. Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of EMNLP*. pages 1520–1530. http://aclweb.org/anthology/D15-1176.

Fei Liu and Julien Perez. 2017. Gated end-to-end memory networks. In *Proceedings of EACL (to appear)*. http://arxiv.org/abs/1610.04211.

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. http://www.aclweb.org/anthology/P15-1145.

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 Natural Language Processing*. pages 151–159. http://www.aclweb.org/anthology/W15-1521.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of CoNLL*. pages 104–113. http://www.aclweb.org/anthology/W13-3512.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. Context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of CoNLL*. pages 51–61. http://aclweb.org/anthology/K/K16/K16-1006.pdf.

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

Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In *Proceedings of NIPS*. pages 2265–2273.

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. http://aclweb.org/anthology/P/P15/P15-2130.pdf.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Tsung-Hsien Wen, and Steve Young. 2016. Neural Belief Tracker: Data-driven dialogue state tracking. In *arXiv preprint: 1606.03777*. http://arxiv.org/abs/1606.03777.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Maria Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of NAACL-HLT*. http://aclweb.org/anthology/N/N16/N16-1018.pdf.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of ICML*. pages 807–814. http://www.icml2010.org/papers/432.pdf.

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. https://doi.org/10.1016/j.artint.2012.07.001.

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. http://anthology.aclweb.org/P16-2074.

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

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of ACL*. pages 425–430. http://www.aclweb.org/anthology/P15-2070.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*. pages 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Siyu Qiu, Qing Cui, Jiang Bian, Bin Gao, and Tie-Yan Liu. 2014. Co-learning of word representations and morpheme representations. In *Proceedings of COLING*. pages 141–150. http://www.aclweb.org/anthology/C14-1015.

Patrick Schone and Daniel Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. In *Proceedings of NAACL*. http://aclweb.org/anthology/N/N01/N01-1024.

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. http://www.aclweb.org/anthology/K15-1026.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2016. Symmetric patterns and coordinations: Fast and enhanced representations of verbs and adjectives. In *Proceedings of NAACL-HLT*. pages 499–505. http://www.aclweb.org/anthology/N16-1060.

Radu Soricut and Franz Och. 2015. Unsupervised morphology induction using word embeddings. In *Proceedings of NAACL-HLT*. pages 1627–1637. http://www.aclweb.org/anthology/N15-1186.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *Proceedings of ACL*. pages 674–680. http://www.aclweb.org/anthology/P15-2111.

Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kuebler, Yannick Versley, Marie Candito, Jennifer Foster, Ines Rehbein, and Lamia Tounsi. 2010. Statistical parsing of morphologically rich languages (SPMRL) What, how and whither. In *Proceedings of the NAACL Workshop on Statistical Parsing of Morphologically-Rich Languages*. pages 1–12. http://www.aclweb.org/anthology/W10-1401.

Joseph P. 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. http://www.aclweb.org/anthology/P10-1040.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: vector space models of semantics. *Journal of Artifical Intelligence Research* 37(1):141–188. https://doi.org/10.1613/jair.2934.

Lyan Verwimp, Joris Pelemans, Hugo Van hamme, and Patrick Wambacq. 2017. Character-word LSTM language models. In *Proceedings of EACL (to appear)*.

Ivan Vulić and Anna Korhonen. 2016a. Is "universal syntax" universally useful for learning distributed word representations? In *Proceedings of ACL*. pages 518–524. http://anthology.aclweb.org/P16-2084.

Ivan Vulić and Anna Korhonen. 2016b. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of ACL*. pages 247–257. http://www.aclweb.org/anthology/P16-1024.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of AAAI*. pages 1112–1119.

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 (to appear)*. http://arxiv.org/abs/1604.04562.

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. https://aclweb.org/anthology/D16-1157.

Jason D. Williams, Antoine Raux, and Matthew Henderson. 2016. The Dialog State Tracking Challenge series: A review. *Dialogue & Discourse* 7(3):4–33. http://dad.uni-bielefeld.de/index.php/dad/article/view/3685.

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. https://doi.org/10.1145/2661829.2662038.

Steve Young. 2010. Cognitive User Interfaces. *IEEE Signal Processing Magazine* .

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of ACL*. pages 545–550. http://www.aclweb.org/anthology/P14-2089.

Britta Zeller, Jan Šnajder, and Sebastian Padó. 2013. DErivBase: Inducing and evaluating a derivational morphology resource for German. In *Proceedings of ACL*. pages 1201–1211. http://www.aclweb.org/anthology/P13-1118.

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. http://www.aclweb.org/anthology/D13-1141.