On the Role of Seed Lexicons in Learning Bilingual Word Embeddings

Ivan Vulić and **Anna Korhonen**
Language Technology Lab
DTAL, University of Cambridge
{iv250, alk23}@cam.ac.uk

Abstract

A shared bilingual word embedding space (SBWES) is an indispensable resource in a variety of cross-language NLP and IR tasks. A common approach to the SBWES induction is to learn a mapping function between monolingual semantic spaces, where the mapping critically relies on a seed word lexicon used in the learning process. In this work, we analyze the importance and properties of seed lexicons for the SBWES induction across different dimensions (i.e., lexicon source, lexicon size, translation method, translation pair reliability). On the basis of our analysis, we propose a simple but effective hybrid bilingual word embedding (BWE) model. This model (HYBWE) learns the mapping between two monolingual embedding spaces using only highly reliable symmetric translation pairs from a seed document-level embedding space. We perform bilingual lexicon learning (BLL) with 3 language pairs and show that by carefully selecting reliable translation pairs our new HYBWE model outperforms benchmarking BWE learning models, all of which use more expensive bilingual signals. Effectively, we demonstrate that a SBWES may be induced by leveraging only a very weak bilingual signal (document alignments) along with monolingual data.

1 Introduction

Dense real-valued vector representations of words or word embeddings (WEs) have recently gained increasing popularity in natural language processing (NLP), serving as invaluable features in a broad range of NLP tasks, e.g., (Turian et al., 2010; Collobert et al., 2011; Chen and Manning, 2014). Several studies have showcased a direct link and comparable performance to "more traditional" distributional models (Turney and Pantel, 2010). Yet the widely used skip-gram model with negative sampling (SGNS) (Mikolov et al., 2013b) is considered as the state-of-the-art word representation model, due to its simplicity, fast training, as well as its solid and robust performance across a wide variety of semantic tasks (Baroni et al., 2014; Levy and Goldberg, 2014b; Levy et al., 2015).

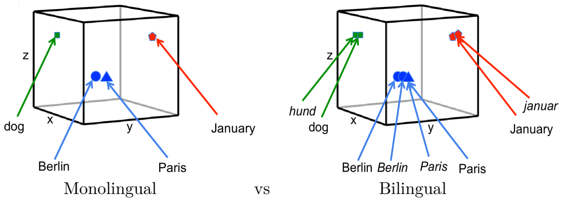


Figure 1: A toy example of a 3-dimensional monolingual vs shared bilingual word embedding space (further SBWES) from Gouws et al. (2015).

Research interest has recently extended to bilingual word embeddings (BWEs). BWE learning models focus on the induction of a *shared bilingual word embedding space* (SBWES) where words from both languages are represented in a uniform language-independent manner such that similar words (regardless of the actual language) have similar representations (see Fig. 1). A variety of BWE learning models have been proposed, differing in the essential requirement of a *bilingual signal* necessary to construct such a SBWES (discussed later in Sect. 2). SBWES may be used to support many tasks, e.g., computing cross-lingual/multilingual semantic word similarity (Faruqui and Dyer, 2014), learning bilingual word lexicons (Mikolov et al., 2013a; Gouws et al., 2015; Vulić et al., 2016), cross-lingual entity linking (Tsai and Roth, 2016), parsing (Guo et al., 2015; Johannsen et al., 2015), machine translation (Zou et al., 2013), or cross-lingual information retrieval (Vulić and Moens, 2015; Mitra et al., 2016).

BWE models should have two desirable properties: **(P1)** leverage (large) monolingual training sets tied together through a bilingual signal, **(P2)** use as inexpensive bilingual signal as possible in order to learn a SBWES in a scalable and widely applicable manner across languages and domains.

While we provide a classification of related work, that is, different BWE models according to these properties in Sect. 2.1, the focus of this work is on a popular class of models labeled *Post-Hoc Mapping with Seed Lexicons*. These models operate as follows (Mikolov et al., 2013a; Dinu et al., 2015; Lazaridou et al., 2015; Ammar et al., 2016): (1) two separate non-aligned monolingual embedding spaces are induced using any monolingual WE learning model (SGNS is the typical choice), (2) given a *seed lexicon* of word translation pairs as the bilingual signal for training, a mapping function is learned which ties the two monolingual spaces together into a SBWES.

All existing work on this class of models assumes that high-quality training seed lexicons are readily available. In reality, little is understood regarding what constitutes a high quality seed lexicon, even with "traditional" distributional models (Gaussier et al., 2004; Holmlund et al., 2005; Vulić and Moens, 2013). Therefore, in this work we ask *whether BWE learning could be improved by making more intelligent choices when deciding over seed lexicon entries*. In order to do this we delve deeper into the cross-lingual mapping problem by analyzing a spectrum of seed lexicons with respect to controllable parameters such as lexicon source, its size, translation method, and translation pair reliability.

The contributions of this paper are as follows:

(C1) We present a systematic study on the importance of seed lexicons for learning mapping functions between monolingual WE spaces.

(C2) Given the insights gained, we propose a simple yet effective hybrid BWE model HYBWE that removes the need for readily available seed lexicons, and satisfies properties P1 and P2. HYBWE relies on an inexpensive seed lexicon of highly reliable word translation pairs obtained by a document-level BWE model (Vulić and Moens, 2016) from document-aligned comparable data.

(C3) Using a careful pair selection process when constructing a seed lexicon, we show that in the BLL task HYBWE outperforms a BWE model of Mikolov et al. (2013a) which relies on readily available seed lexicons. HYBWE also outperforms state-of-the-art models of (Hermann and Blunsom, 2014b; Gouws et al., 2015) which require sentence-aligned parallel data.

2 Learning SBWES using Seed Lexicons

Given source and target language vocabularies V^S and V^T, all BWE models learn a representation of each word $w \in V^S \sqcup V^T$ in a SBWES as a real-valued vector: $\mathbf{w} = [f_1, \ldots, f_d]$, where $f_k \in \mathbb{R}$ denotes the value for the k-th cross-lingual feature for w within a d-dimensional SBWES. Semantic similarity $sim(w, v)$ between two words $w, v \in V^S \sqcup V^T$ is then computed by applying a similarity function (SF), e.g. cosine (*cos*) on their representations in the SBWES: $sim(w, v) = SF(\mathbf{w}, \mathbf{v}) = cos(\mathbf{w}, \mathbf{v})$.

2.1 Related Work: BWE Models and Bilingual Signals

BWE models may be clustered into four different types according to bilingual signals used in training, and properties P1 and P2 (see Sect. 1). Upadhyay et al. (2016) provide a similar overview of recent bilingual embedding learning architectures regarding different bilingual signals required for the embedding induction.

(Type 1) *Parallel-Only*: This group of BWE models relies on sentence-aligned and/or word-aligned parallel data as the only data source (Zou et al., 2013; Hermann and Blunsom, 2014a; Kočiský et al., 2014; Hermann and Blunsom, 2014b; Chandar et al., 2014). In addition to an expensive bilingual signal (colliding with P2), these models do not leverage larger monolingual datasets for training (not satisfying P1).

(Type 2) *Joint Bilingual Training*: These models jointly optimize two monolingual objectives, with the cross-lingual objective acting as a cross-lingual regularizer during training (Klementiev et al., 2012; Gouws et al., 2015; Soyer et al., 2015; Shi et al., 2015; Coulmance et al., 2015). The idea may be summarized by the simplified formulation (Luong et al., 2015): $\gamma(Mono_S + Mono_T) + \delta Bi$. The monolingual objectives $Mono_S$ and $Mono_T$ ensure that similar words in each language are assigned similar embeddings and aim to capture the semantic structure of each language, whereas the cross-lingual objective Bi ensures that similar words across languages are assigned similar embeddings. It ties the two monolingual spaces together into a SBWES (thus satisfying P1). Parameters γ and δ govern the influence of the monolingual and bilingual components.[1] The main disadvantage of Type 2 models is the costly parallel data needed for the bilingual signal (thus colliding with P2).

(Type 3) *Pseudo-Bilingual Training*: This set of models requires *document alignments* as bilingual signal to induce a SBWES. Vulić and Moens (2016) create a collection of pseudo-bilingual documents by merging every pair of aligned documents in training data, in a way that preserves important local information: words that appeared next to other words within the same language and those that appeared in the same region of the document across different languages. This collection is then used to train word embeddings with monolingual SGNS from `word2vec`.

With pseudo-bilingual documents, the "context" of a word is redefined as a mixture of neighbouring words (in the original language) and words that appeared in the same region of the document (in the ”foreign” language). The bilingual contexts for each word in each document steer the final model towards constructing a SBWES. The advantage over other BWE model types lies in exploiting weaker document-level bilingual signals (satisfying P2), but these models are unable to exploit monolingual corpora during training (unlike Type 2 or Type 4; thus colliding with P1).

(Type 4) *Post-Hoc Mapping with Seed Lexicons*: These models learn post-hoc mapping functions between monolingual WE spaces induced separately for two different languages (e.g., by SGNS). All Type 4 models (Mikolov et al., 2013a; Faruqui and Dyer, 2014; Dinu et al., 2015; Lazaridou et al., 2015) rely on readily available seed lexicons of highly frequent words obtained by e.g. *Google Translate* (GT) to learn the mapping (again colliding with P2), but they are able to satisfy P1.

2.2 Post-Hoc Mapping with Seed Lexicons: Methodology and Lexicons

Key Intuition One may infer that a type-*hybrid* procedure which would retain only highly reliable translation pairs obtained by a Type 3 model as a seed lexicon for Type 4 models effectively satisfies both requirements: (P1) unlike Type 1 and Type 3, it can learn from monolingual data and tie two monolingual spaces using the highly reliable translation pairs, (P2) unlike Type 1 and Type 2, it does not require parallel data; unlike Type 4, it does not require external lexicons and translation systems. The only bilingual signal required are document alignments. Therefore, our focus is on novel less expensive Type 4 models.

Overview The standard learning setup we use is as follows: First, two monolingual embedding spaces, $\mathbb{R}^{d_S}$ and $\mathbb{R}^{d_T}$, are induced separately in each of the two languages using a standard monolingual WE model such as CBOW or SGNS. d_S and d_T denote the dimensionality of monolingual WE spaces. The bilingual signal is a seed lexicon, i.e., a list of word translation pairs (x_i, y_i), where $x_i \in V^S$, $y_i \in V^T$, and $\mathbf{x_i} \in \mathbb{R}^{d_S}$, $\mathbf{y_i} \in \mathbb{R}^{d_T}$.

Learning Objectives Training is cast as a multivariate regression problem: it implies learning a function that maps the source language vectors from the training data to their corresponding target language vectors. A standard approach (Mikolov et al., 2013a; Dinu et al., 2015) is to assume a linear map $\mathbf{W} \in \mathbb{R}^{d_S \times d_T}$, where a L_2-regularized least-squares error objective (i.e., ridge regression) is used to learn the map $\mathbf{W}$. The map is learned by solving the following optimization problem (typically by stochastic gradient descent (SGD)):

$$\min_{\mathbf{W} \in \mathbb{R}^{d_S \times d_T}} ||\mathbf{XW} - \mathbf{Y}||_F^2 + \lambda ||\mathbf{W}||_F^2 \quad (1)$$

$\mathbf{X}$ and $\mathbf{Y}$ are matrices obtained through the respective concatenation of source language and target language vectors from training pairs. Once the linear map $\mathbf{W}$ is estimated, any previously unseen source language word vector $\mathbf{x_u}$ may be straightforwardly mapped into the target language embedding space $\mathbb{R}^{d_T}$ as $\mathbf{Wx_u}$. After mapping all vectors $\mathbf{x}$, $x \in V^S$, the target embedding space $\mathbb{R}^{d_T}$ in fact serves as SBWES.[2]

[1] Type 1 models may be considered a special case of Type 2 models: Setting $\gamma = 0$ reduces Type 2 models to Type 1 models trained solely on parallel data, e.g., (Hermann and Blunsom, 2014b; Chandar et al., 2014). $\gamma = 1$ results in the models from (Klementiev et al., 2012; Gouws et al., 2015; Soyer et al., 2015; Coulmance et al., 2015).

[2] Another possible objective (found in the zero-shot learning literature) is a margin-based ranking loss (Weston et al., 2011; Lazaridou et al., 2015). We omit the results with this objective for brevity, and due to the fact that similar trends are observed as with (more standard) linear maps.

Seed Lexicon Source and Translation Method Prior work on post-hoc mapping with seed lexicons used a translation system (i.e., GT) to translate highly frequent English words to other languages such as Czech, Spanish (Mikolov et al., 2013a; Gouws et al., 2015) or Italian (Dinu et al., 2015; Lazaridou et al., 2015). This method presupposes the availability and high quality of such an external translation system. To simulate this setup, we take as a starting point the BNC word frequency list from Kilgarriff (1997) containing $6,318$ most frequent English lemmas. The list is then translated to other languages via GT. We call the BNC-based lexicons obtained by employing Google Translate BNC+GT.

In this paper, we propose another option: first, we learn the "first" SBWES (i.e., SBWES-1) using another BWE model (see Sect. 2.1), and then translate the BNC list through SBWES-1 by retaining the nearest cross-lingual neighbor $y_i \in V^T$ for each x_i in the BNC list which is represented in SBWES-1. The pairs (x_i, y_i) constitute the seed lexicon needed for learning the mapping between monolingual spaces, that is, to induce the final SBWES-2.

Although in theory any BWE induction model may be used to induce SBWES-1, we rely on a document-level Type 3 BWE induction model from (Vulić and Moens, 2016), since it requires only document alignments as (weak) bilingual signal. The resulting hybrid BWE induction model (HYBWE) combines the output of a Type 3 model (SBWES-1) and a Type 4 model (SBWES-2). This seed lexicon and BWE learning variant is called BNC+HYB.

Our new hybrid model allows us to also use source language words occurring in SBWES-1 sorted by frequency as seed lexicon source, again leaning on the intuition that higher frequency phenomena are more reliably translated using statistical models. Their translations can also be found through SBWES-1 to obtain seed lexicon pairs (x_i, y_i). This variant is called HFQ+HYB.

Another possibility, recently introduced by Kiros et al. (2015) for vocabulary expansion in monolingual settings, relies on all words shared between two vocabularies to learn the mapping. In this work, we test the ability and limits of such orthographic evidence in cross-lingual settings: seed lexicon pairs are (x_i, x_i), where $x_i \in V^S$ and $x_i \in V^T$. This seed lexicon variant is called ORTHO.

Seed Lexicon Size While all prior reported only results with restricted seed lexicon sizes only (i.e., 1K, 2K and 5K lexicon pairs are used as standard), in this work we provide a full-fledged analysis of the influence of seed lexicon size on the SBWES performance in cross-lingual tasks. More extreme settings are also investigated, in the attempt to answer two important questions: (1) Can a Type 4 SBWES be induced in a limited setting with only a few hundred lexicon pairs available (e.g., 100-500)? (2) Can the Type 4 models profit from the inclusion of more seed lexicon pairs (e.g., more than 5K, even up to 40K-50K lexicon pairs)?

Translation Pair Reliability When building seed lexicons through SBWES-1 (i.e., BNC+HYB and HFQ+HYB methods), it is possible to control for the reliability of translation pairs to be included in the final lexicon, with the idea that the use of only highly reliable pairs can potentially lead to an improved SBWES-2. A simple yet effective reliability reliability feature for translation pairs is the *symmetry constraint* (Peirsman and Padó, 2010; Vulić and Moens, 2013) : two words $x_i \in V^S$ and $y_i \in V^S$ are used as seed lexicon pairs only if they are mutual nearest neighbours given their representations in SBWES-1. The two variants of seed lexicons with only symmetric pairs are BNC+HYB+SYM and HFREQ+HYB+SYM. We also test the variants without the symmetry constraint (i.e., BNC+HYB+ASYM and HFQ+HYB+ASYM).

Even more conservative reliability measures may be applied by exploiting the scores in the lists of translation candidates ranked by their similarity to the cue word x_i. We investigate a symmetry constraint with a *threshold*: two words $x_i \in V^S$ and $y_i \in V^S$ are included as seed lexicon pair (x_i, y_i) iff they are mutual nearest neighbours in SBWES-1 and it holds:

$$sim(x_i, y_i) - sim(x_i, z_i) > THR \quad (2)$$

$$sim(y_i, x_i) - sim(y_i, w_i) > THR \quad (3)$$

where $z_i \in V^T$ is the second best translation candidate for x_i, and $w_i \in V^S$ for y_i. THR is a parameter which specifies the margin between the two best translation candidates. The intuition is that highly unambiguous and monosemous translation pairs (which is reflected in higher score margins) are also highly reliable.[3]

[3] Other (more elaborate) reliability measures exist in the

3 Experimental Setup

Task: Bilingual Lexicon Learning (BLL) After the final SBWES is induced, given a list of n source language words $x_{u1}, \ldots, x_{un}$, the task is to find a target language word t for each x_u in the list using the SBWES. t is the target language word closest to the source language word x_u in the induced SBWES, also known as the *cross-lingual nearest neighbor*. The set of learned n (x_u, t) pairs is then run against a gold standard BLL test set. Following the standard practice (Mikolov et al., 2013a; Dinu et al., 2015), for all Type 4 models, all pairs containing any of the test words $x_{u1}, \ldots, x_{un}$ are removed from training seed lexicons.

Test Sets For each language pair, we evaluate on standard 1,000 ground truth one-to-one translation pairs built for three language pairs: Spanish (ES)-, Dutch (NL)-, Italian (IT)-English (EN) by Vulić and Moens (2013). The dataset is generally considered a benchmarking test set for BLL models that learn from non-parallel data, and is available online.[4] We have also experimented with two other benchmarking BLL test sets (Bergsma and Durme, 2011; Leviant and Reichart, 2015) observing a very similar relative performance of all the models in our comparison.

Evaluation Metrics We measure the BLL performance using the standard *Top 1* accuracy (Acc_1) metric (Gaussier et al., 2004; Mikolov et al., 2013a; Gouws et al., 2015).[5]

Baseline Models To induce SBWES-1, we resort to document-level embeddings of Vulić and Moens (2016) (Type 3). We also compare to results obtained directly by their model (BWESG) to measure the performance gains with HYBWE.

To compare with a representative Type 2 model, we opt for the BilBOWA model of Gouws et al. (2015) due to its solid performance and robustness in the BLL task when trained on general-domain corpora such as Wikipedia (Luong et al., 2015), its reduced complexity reflected in fast computations on massive datasets, as well as its public availability.[6] In short, BilBOWA combines the adapted SGNS for monolingual objectives together with a cross-lingual objective that minimizes the L_2-loss between the bag-of-word vectors of parallel sentences. BilBOWA uses the same training setup as HYBWE (monolingual datasets plus a bilingual signal), but relies on a stronger bilingual signal (sentence alignments as opposed to HYBWE's document alignments).

We also compare with a benchmarking Type 1 model from sentence-aligned parallel data called BiCVM (Hermann and Blunsom, 2014b). Finally, a SGNS-based BWE model with the BNC+GT seed lexicon is taken as a baseline Type 4 model (Mikolov et al., 2013a).[7]

Training Data and Setup We use standard training data and suggested settings to obtain BWEs for all models involved in comparison. We retain the 100K most frequent words in each language for all models. To induce monolingual WE spaces, two monolingual SGNS models were trained on the cleaned and tokenized Wikipedias from the Polyglot website (Al-Rfou et al., 2013) using SGD with a global learning rate of 0.025. For BilBOWA, as in the original work (Gouws et al., 2015), the bilingual signal for the cross-lingual regularization is provided by the first 500K sentences from Europarl.v7 (Tiedemann, 2012). We use SGD with a global rate of 0.15.[8] The window size is varied from 2 to 16 in steps of 2, and the best scoring model is always reported in all comparisons.

BWESG was trained on the cleaned and tokenized document-aligned Wikipedias available online[9], SGD on pseudo-bilingual documents with a global rate 0.025. For BiCVM, we use the tool released by its authors[10] and train on the whole Europarl.v7 for each language pair: we train an additive model, with hinge loss margin set to d (i.e., dimensionality) as in the original paper, batch size of 50, and noise parameter of 10. All BiCVM models are trained with 200 iterations.

For all models, we obtain BWEs with $d = 40, 64, 300, 500$, but we report only results with 300-dimensional BWEs as similar trends were observed with other d-s. Other parameters are: 15 epochs, 15 negatives, subsampling rate $1e-4$.

literature (Smith and Eisner, 2007; Tu and Honavar, 2012; Vulić and Moens, 2013), but we do not observe any significant gains when resorting to the more complex reliability estimates.

[4] http://people.cs.kuleuven.be/~ivan.vulic/

[5] Similar trends are observed within a more lenient setting with Acc_5 and Acc_{10} scores, but we omit these results for clarity and the fact that the actual BLL performance is best reflected in Acc_1 scores (i.e., best translation only).

[6] https://github.com/gouwsmeister/bilbowa

[7] For details concerning all baseline models, the reader is encouraged to check the relevant literature.

[8] Suggested by the authors (personal correspondence).

[9] http://linguatools.org/tools/corpora/

[10] https://github.com/karlmoritz/bicvm

BNC+GT	**BNC+HYB+ASYM**	**BNC+HYB+SYM**	**HFQ+HYB+ASYM**	**HFQ+HYB+SYM**	**ORTHO**
casamiento	*casamiento*	*casamiento*	*casamiento*	*casamiento*	*casamiento*
marriage	marry	*marriage*	*marriage*	*marriage*	maría
marry	*marriage*	marry	marry	marry	señor
marrying	marrying	marrying	betrothal	betrothal	doña
betrothal	wed	wedding	marrying	marrying	juana
wedding	wedding	betrothal	wedding	wedding	noche
wed	betrothal	wed	daughter	wed	amor
elopement	remarry	marriages	betrothed	elopement	guerra

Table 1: Nearest EN neighbours of the Spanish word *casamiento (marriage)* with different seed lexicons.

Model	**ES-EN**	**NL-EN**	**IT-EN**
BiCVM (Type 1)	0.532	0.583	0.569
BilBOWA (Type 2)	0.632	0.636	0.647
BWESG (Type 3)	0.676	0.626	0.643
BNC+GT (Type 4)	0.677	0.641	0.646
ORTHO	0.233	0.506	0.224
BNC+HYB+ASYM	0.673	0.626	0.644
BNC+HYB+SYM (3388; 2738; 3145)	0.681	**0.658***	0.663*
HFQ+HYB+ASYM	0.673	0.596	0.635
HFQ+HYB+SYM	**0.695***	0.657*	**0.667***

Table 2: Acc_1 scores in a standard BLL setup (for Type 4 models): all seed lexicons contain 5K translation pairs, except for BNC+HYB+SYM (its sizes provided in parentheses). * denotes a statistically significant improvement over baselines and BNC+GT using McNemar's statistical significance test with the Bonferroni correction, $p < 0.05$.

4 Results and Discussion

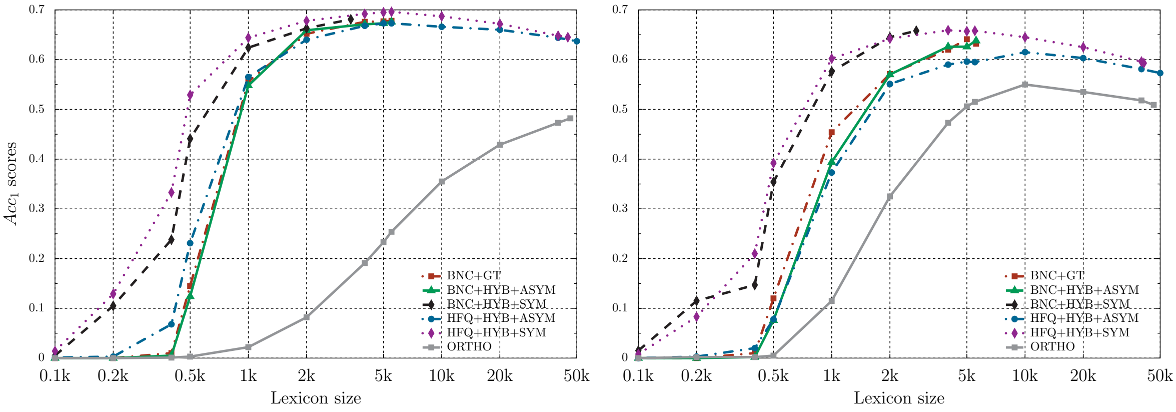
Exp. I: Standard BLL Setting First, we replicate the previous BLL setups with Type 4 models from (Mikolov et al., 2013a; Dinu et al., 2015) by relying on seed lexicons of exactly 5K word pairs (except for BNC+HYB+SYM which exhausts all possible pairs before the 5K limit) sorted by frequency of the source language word. Results with different lexicons for the three language pairs are summarized in Table 2, while Table 1 shows examples of nearest neighbour words for a Spanish word not present in any of the training lexicons.

Table 1 provides evidence for our first insight: Type 4 models do not necessarily require external lexicons (such as the BNC+GT model) to learn a semantically plausible SBWES (i.e., the lists of nearest neighbours are similar for all lexicons excluding ORTHO). Table 1 also suggests that the choice of seed lexicon pairs may strongly influence the properties of the resulting SBWES. Due to its design, ORTHO finds a mapping which naturally brings foreign words appearing in the English vocabulary closer in the induced SBWES.

This first batch of quantitative results already shows that Type 4 models with inexpensive automatically induced lexicons (i.e., HYBWE) are on a par with or even better than Type 4 models relying on external resources or translation systems. In addition, the best reported scores using the more constrained symmetric BNC/HFQ+HYB+SYM lexicon variants are higher than those for three baseline models (of Type 1, Type 2, and Type 3) that previously held highest scores on the BLL test sets (Vulić and Moens, 2016). These improvements over the baseline models and BNC+GT are statistically significant (using McNemar's statistical significance test, $p < 0.05$). Table 2 also suggests that a careful selection of reliable pairs can lead to peak performances even with a lower number of pairs, i.e., see the results of BNC+HYB+SYM.

Exp. II: Lexicon Size BLL results for ES-EN and NL-EN obtained by varying the seed lexicon sizes are displayed in Fig. 2(a) and 2(b). Results for IT-EN closely follow the patterns observed with ES-EN. BNC+HYB+SYM and HFQ+HYB+ASYM – the two models that do not blindly use all potential training pairs, but rely on sets of symmetric pairs (i.e., they include the simple measure of translation pair reliability) – display the best performance across all lexicon sizes. The finding confirms the intuition that a more intelligent pair selection strategy is essential for Type 4 BWE models. HFQ+HYB+SYM – a simple hybrid BWE model (HYBWE) combining a document-level Type 3 model with a Type 4 model and translation reliability detection – is the strongest BWE model overall (see also Table 2 again).

HYBWE-based models which do not perform any pair selection (i.e., BNC/HFQ+HYB+ASYM) closely follow the behaviour of the GT-based model. This demonstrates that an external lexicon or translation system may be safely replaced



(a) Spanish-English

(b) Dutch-English

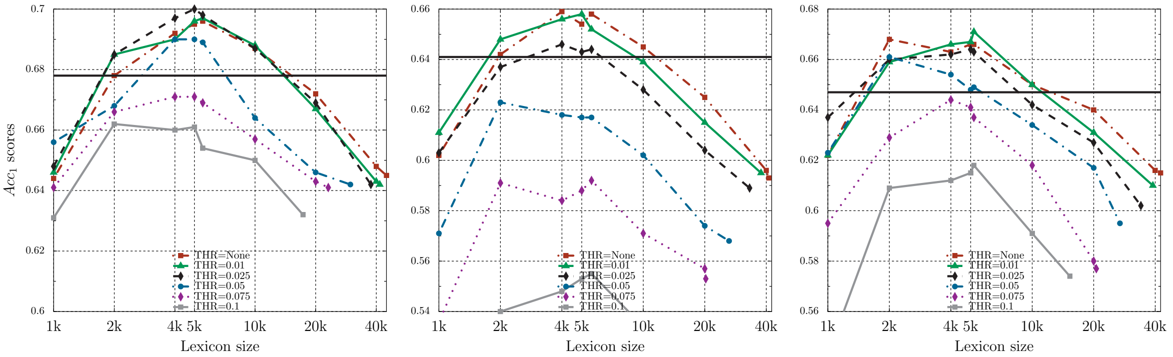
Figure 2: BLL results (Acc_1) across different seed lexicon sizes for all lexicons. x axes are in log scale.

by a document-level embedding model without any significant performance loss in the BLL task. The ORTHO-based model falls short of its competitors. However, we observe that even this model with the learning setting relying on the cheapest bilingual signal may lead to reasonable BLL scores, especially for the more related NL-EN pair.

The two models with the symmetry constraint display a particularly strong performance with settings relying on scarce resources (i.e., only a small portion of training pairs is available). For instance, HFQ+HYB+SYM scores 0.129 for ES-EN with only 200 training pairs (vs 0.002 with BNC+GT), and 0.529 with 500 pairs (vs 0.145 with BNC+GT). On the other hand, adding more pairs does not lead to an improved BLL performance. In fact, we observe a slow and steady decrease in performance with lexicons containing $10,000$ and more training pairs for all HYBWE variants. The phenomenon may be attributed to the fact that highly frequent words receive more accurate representations in SBWES-1, and adding less frequent and, consequently, less accurate training pairs to the SBWES-2 learning process brings in additional noise. In plain language, when it comes to seed lexicons Type 4 models prefer quality over quantity.

Exp. III: Translation Pair Reliability In the next experiment, we vary the threshold value THR (see sect. 2.2) in the HFQ+HYB+SYM variant with the following values in comparison: 0.0 (None), $0.01, 0.025, 0.05, 0.075, 0.1$. We investigate whether retaining only highly unambiguous pairs would lead to even better BLL performance. The results for all three language pairs are summarized in Fig. 3(a)-3(c). The results for all variant models again decrease when employing larger lexicons (due to the usage of less frequent word pairs in training). We observe that a slightly stricter selection criterion (i.e., THR $=$ $0.01, 0.025$) also leads to slightly improved peak BLL scores for ES-EN and IT-EN around the 5K region. The improvements, however, are not statistically significant. On the other hand, a too conservative pair selection criterion with higher threshold values significantly deteriorates the overall performance of HYBWE with HFQ+HYB+SYM. The conservative criteria discard plenty of potentially useful training pairs. Therefore, as one line of future research, we plan to investigate more sophisticated models for the selection of reliable seed lexicon pairs that will lead to a better trade-off between the lexicon size and reliability of the pairs.

Exp. IV: Another Task - Suggesting Word Translations in Context (SWTC) In the final experiment, we test whether the findings originating from the BLL task generalize to another cross-lingual semantic task: *suggesting word translations in context* (SWTC) recently proposed by Vulić and Moens (2014). Given an occurrence of a polysemous word $w \in V^S$, the SWTC task is to choose the correct translation in the target language of that particular occurrence of w from the given set $\mathcal{TC}(w) = \{t_1, \ldots, t_{tq}\}$, $\mathcal{TC}(w) \subseteq V^T$, of its tq possible translations/meanings. Whereas in the BLL task the candidate search is performed over the entire vocabulary V^T, the set $TC(w)$ typically comprises only a few pre-selected words/senses. One may refer to $\mathcal{TC}(w)$ as an inventory of translation candidates for w. The best scoring translation candidate in the ranked list is then the correct translation for that particular occurrence of w observing its local context $Con(w)$. SWTC is an extended



(a) Spanish-English

(b) Dutch-English

(c) Italian-English

Figure 3: BLL results across different threshold (THR) values with the HFQ+HYB+SYM seed lexicons. Higher thresholds imply less ambiguous word translation pairs. Thicker horizontal lines denote the best score from any of the baseline models. x axes are in log scale.

Model	**ES-EN**	**NL-EN**	**IT-EN**
NO CONTEXT	0.406	0.433	0.408
BEST SYSTEM (Vulić and Moens, 2014)	0.703	0.712	0.789
BICVM (TYPE 1)	0.506	0.586	0.522
BILBOWA (TYPE 2)	0.586	0.656	0.589
BWESG (TYPE 3)	0.783	0.858	0.792
BNC+GT (TYPE 4)	0.794	0.858	0.783
ORTHO	0.647	0.794	0.678
BNC+HYB+ASYM	0.806*	0.872	0.778
BNC+HYB+SYM (3839; 3117; 3693)	**0.808***	**0.875***	**0.814***
HFQ+HYB+ASYM	0.789	0.864	0.781
HFQ+HYB+SYM (THR = None)	0.792	0.869	0.786
HFQ+HYB+SYM (THR=0.01)	0.792	0.858	0.789
HFQ+HYB+SYM (THR=0.025)	0.800	0.853	0.792

Table 3: Acc_1 scores in the SWTC task. All seed lexicons contain 6K translation pairs, except for BNC+HYB+SYM (its sizes provided in parentheses). * denotes a statistically significant improvement over baselines and BNC+GT using McNemar's statistical significance test with the Bonferroni correction, $p < 0.05$.

cross-lingual variant of the task proposed by Huang et al. (2012) which evaluates monolingual context-sensitive semantic similarity of words in sentential context, and it is also very related to cross-lingual lexical substitution (Mihalcea et al., 2010).

To isolate the performance of each BWE induction model from the details of the SWTC setup, we use the same approach with all models: we opt for the SWTC framework proven to yield excellent results with BWEs in the SWTC task (Vulić and Moens, 2016). In short, the context bag $Con(w) = \{cw_1, \ldots, cw_r\}$ is obtained by harvesting all r words that occur with w in the sentence. The vector representation of $Con(w)$ is the d-dimensional embedding computed by aggregating over all word embeddings for each $cw_j \in Con(w)$ using standard *addition* as the compositional operator (Mitchell and Lapata, 2008) which was proven a robust choice (Milajevs et al., 2014):

$$\mathbf{Con(w)} = \mathbf{cw}_1 + \mathbf{cw}_2 + \ldots + \mathbf{cw}_r \quad (4)$$

where $\mathbf{cw}_j$ is the embedding of the j-th context word, and $\mathbf{Con(w)}$ is the resulting embedding of the context bag $Con(w)$. Finally, for each $t_j \in \mathcal{TC}(w)$, the context-sensitive similarity with w is computed as: $sim(w, t_j, Con(w)) = cos(\mathbf{Con(w)}, \mathbf{t}_j)$, where $\mathbf{Con(w)}$ and $\mathbf{t}_j$ are representations of the (sentential) context bag and the candidate translation t_j in the same SBWES.[11]

The evaluation set consists of 360 sentences for 15 polysemous nouns (24 sentences for each noun) in each of the three languages: Spanish, Dutch, Italian, along with the single gold standard single word English translation given the sentential context.[12] Table 3 summarizes the results (Acc_1 scores) in the SWTC task. NO-CONTEXT refers to the context-insensitive majority baseline obtained by BNC+GT (i.e., it always chooses the most semantically similar translation candidate at the word type level). We also report the results of the best SWTC model from Vulić and Moens (2014).

The results largely support the claims established with the BLL evaluation. An exter-

[11] The same ranking of different models (with lower absolute scores) is observed when adapting the monolingual lexical substitution framework of Melamud et al. (2015) to the SWTC task as done by Vulić and Moens (2016).

[12] The SWTC evaluation set is available online at: http://aclweb.org/anthology/attachments/D/D14/D14-1040.Attachment.zip

nal seed lexicon of BNC+GT may be safely replaced by an automatically induced inexpensive seed lexicon (as in HYBWE with BNC+HYB+SYM/ASYM). The best performing models are again BNC+HYB+SYM and HFQ+HYB+SYM. The comparison of ASYM and SYM lexicon variants further suggests that filtering translation pairs using the symmetry constraint again leads to consistent improvements, but stricter selection criteria with higher thresholds do not lead to significant performance boosts, and may even hurt the performance (see the results for NL-EN). Various HYBWE variants significantly improve over baseline BWE models (Types 1-4), also outperforming previous best SWTC results.

5 Conclusions and Future Work

We presented a detailed analysis of the importance and properties of seed bilingual lexicons in learning bilingual word embeddings (BWEs) which are valuable for many cross-lingual/multilingual NLP tasks. On the basis of the analysis, we proposed a simple yet effective hybrid bilingual word embedding model called HYBWE. It learns the mapping between two monolingual embedding spaces using only highly reliable symmetric translation pairs from an inexpensive seed document-level embedding space. The results in the tasks of (1) bilingual lexicon learning and (2) suggesting word translations in context demonstrate that – due to its careful selection of reliable translation pairs for seed lexicons – HYBWE outperforms benchmarking BWE induction models, all of which use more expensive bilingual signals for training.

In future work, we plan to investigate other methods for seed pairs selection, settings with scarce resources (Agić et al., 2015; Zhang et al., 2016), other context types inspired by recent work in the monolingual settings (Levy and Goldberg, 2014a; Melamud et al., 2016), as well as model adaptations that can work with multi-word expressions. Encouraged by the excellent results, we also plan to test the portability of the approach to more language pairs, and other tasks and applications.

Acknowledgments

This work is supported by ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (no 648909). The authors are grateful to Roi Reichart and the anonymous reviewers for their helpful comments and suggestions.

References

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *ACL*, pages 268–272.

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

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR*, abs/1602.01925.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*, pages 238–247.

Shane Bergsma and Benjamin Van Durme. 2011. Learning bilingual lexicons using the visual similarity of labeled web images. In *IJCAI*, pages 1764–1769.

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

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

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.

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

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

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

Éric Gaussier, Jean-Michel Renders, Irina Matveeva, Cyril Goutte, and Hervé Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *ACL*, pages 526–533.

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

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

Karl Moritz Hermann and Phil Blunsom. 2014a. Multilingual distributed representations without word alignment. In *ICLR*.

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

Jon Holmlund, Magnus Sahlgren, and Jussi Karlgren. 2005. Creating bilingual lexica using reference wordlists for alignment of monolingual semantic vector spaces. In *NODALIDA*, pages 71–77.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *ACL*, pages 873–882.

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

Adam Kilgarriff. 1997. Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2):135–155.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *NIPS*.

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

Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. In *ACL*, pages 224–229.

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

Ira Leviant and Roi Reichart. 2015. Judgment language matters: Multilingual vector space models for judgment language aware lexical semantics. *CoRR*, abs/1508.00106.

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

Omer Levy and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. In *NIPS*, pages 2177–2185.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the ACL*, 3:211–225.

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.

Oren Melamud, Omer Levy, and Ido Dagan. 2015. A simple word embedding model for lexical substitution. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 1–7.

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

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 task 2: Cross-lingual lexical substitution. In *SEMEVAL*, pages 9–14.

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

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

Dmitrijs Milajevs, Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Matthew Purver. 2014. Evaluating neural word representations in tensor-based compositional settings. In *EMNLP*, pages 708–719.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *ACL*, pages 236–244.

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

Yves Peirsman and Sebastian Padó. 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *NAACL*, pages 921–929.

Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. 2015. Learning cross-lingual word embeddings via matrix co-factorization. In *ACL*, pages 567–572.

David A. Smith and Jason Eisner. 2007. Bootstrapping feature-rich dependency parsers with entropic priors. In *EMNLP-CoNLL*, pages 667–677.

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

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *LREC*, pages 2214–2218.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *NAACL-HLT*.

Kewei Tu and Vasant Honavar. 2012. Unambiguity regularization for unsupervised learning of probabilistic grammars. In *EMNLP-CoNLL*, pages 1324–1334.

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

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.

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

Ivan Vulić and Marie-Francine Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *EMNLP*, pages 1613–1624.

Ivan Vulić and Marie-Francine Moens. 2014. Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. In *EMNLP*, pages 349–362.

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

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

Ivan Vulić, Douwe Kiela, Stephen Clark, and Marie-Francine Moens. 2016. Multi-modal representations for improved bilingual lexicon learning. In *ACL*.

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

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag - Multilingual POS tagging via coarse mapping between embeddings. In *NAACL-HLT*.

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