Hierarchical Mapping for Crosslingual Word Embedding Alignment

Ion Madrazo Azpiazu and **Maria Soledad Pera**

Department of Computer Science
Boise State University
{ionmadrazo,solepera}@boisestate.edu

Abstract

The alignment of word embedding spaces in different languages into a common crosslingual space has recently been in vogue. Strategies that do so compute pairwise alignments and then map multiple languages to a single pivot language (most often English). These strategies, however, are biased towards the choice of the pivot language, given that language proximity and the linguistic characteristics of the target language can strongly impact the resultant crosslingual space in detriment of topologically distant languages. We present a strategy that eliminates the need for a pivot language by learning the mappings across languages in a hierarchical way. Experiments demonstrate that our strategy significantly improves vocabulary induction scores in all existing benchmarks, as well as in a new non-English–centered benchmark we built, which we make publicly available.

1 Introduction

Word embeddings have changed how we build text processing applications, given their capabilities for representing the meaning of words (Mikolov et al., 2013a; Pennington et al., 2014; Bojanowski et al., 2017). Traditional embedding-generation strategies create different embeddings for the same word depending on the language. Even if the embeddings themselves are different across languages, their distributions tend to be consistent—the relative distances across word embeddings are preserved regardless of the language (Mikolov et al., 2013b). This behavior has been exploited for crosslingual embedding generation by aligning any two monolingual embeddings spaces into one (Dinu et al., 2014; Xing et al., 2015; Artetxe et al., 2016).

Alignment techniques have been successful in generating bilingual embedding spaces that can later be merged into a crosslingual space using a pivoting language, English being the most common choice. Unfortunately, mapping one language into another suffers from a neutrality problem, as the resultant bilingual space is impacted by language-specific phenomena and corpus-specific biases of the target language (Doval et al., 2018). To address this issue, Doval et al. (2018) propose mapping any two languages into a different *middle* space. This mapping, however, precludes the use of a pivot language for merging multiple bilingual spaces into a crosslingual one, limiting the solution to a bilingual scenario. Additionally, the pivoting strategy suffers from a generalized bias problem, as languages that are the most similar to the pivot obtain a better alignment and are therefore better represented in the crosslingual space. This is because language proximity is a key factor when learning alignments. This is evidenced by the results in Artetxe et al. (2017), which indicate that when using English (Indo-European) as a pivot, the vocabulary induction results for Finnish (Uralic) are about 10 points below the rest of the Indo-European languages under study.

If we want to incorporate *all languages* into the same crosslingual space regardless of their characteristics, we need to go beyond the *train-bilingual/merge-by-pivoting* (**TB/MP**) model, and instead seek solutions that can directly generate crosslingual spaces without requiring a bilingual step. This motivates the design of **HCEG** (Hierarchical Crosslingual Embedding Generation), the hierarchical pivotless approach for generating crosslingual embedding spaces that we present in this paper. HCEG addresses both the language proximity and target-space bias problems by learning a compositional mapping across multiple languages in a hierarchical fashion. This is accomplished by taking advantage of a language family tree for aggregating multiple languages into a single crosslingual space. What distinguishes HCEG from TB/MP strategies is that it does not need to include the pivot language

Transactions of the Association for Computational Linguistics, vol. 8, pp. 361–376, 2020. https://doi.org/10.1162/tacl_a_00320
Action Editor: Eneko Agirre. Submission batch: 7/2019; Revision batch: 2/2020; Published 7/2020.

in all mapping functions. This enables the option to learn mappings between typologically similar languages, known to yield better quality mappings (Artetxe et al., 2017).

The main contributions of our work include:

- A strategy[1] that leverages a *language family tree* for learning mapping matrices that are composed hierarchically to yield crosslingual embedding spaces for language families.

- An analysis of the benefits of hierarchically generating mappings across multiple languages compared to traditional unsupervised and supervised TB/MP alignment strategies.

2 Related Work

Recent interest in crosslingual word embedding generation has led to manifold strategies that can be classified into four groups (Ruder et al., 2017): (1) **Mapping** techniques that rely on a bilingual lexicon for mapping an already trained monolingual space into another (Mikolov et al., 2013b; Artetxe et al., 2017; Doval et al., 2018); (2) **Pseudo-crosslingual** techniques that generate synthetic crosslingual corpora that are then used in a traditional monolingual strategy, by randomly replacing words of a text with their translations (Gouws and Søgaard, 2015; Duong et al., 2016) or by combining texts in various languages into one (Vulić and Moens, 2016); (3) Approaches that only optimize for a **crosslingual objective** function, which require parallel corpora in the form of aligned sentences (Hermann and Blunsom, 2013; Lauly et al., 2014) or texts (Søgaard et al., 2015); and (4) Approaches using a **joint objective** function that optimizes both mono- and cross-lingual loss, that rely on a parallel corpora aligned at the word (Zou et al., 2013; Luong et al., 2015) or sentence level (Gouws et al., 2015; Coulmance et al., 2015).

A key factor for crosslingual embedding generation techniques is the amount of supervised signal needed. Parallel corpora are a scarce resource—even nonexistent for some isolated or low-resource languages. Thus, we focus on mapping-based strategies that can go from requiring just a bilingual lexicon (Mikolov et al., 2013b) to absolutely no supervised signal (Artetxe et al., 2018). This aligns with one of the premises for our research to enable the generation of a single crosslingual embedding space for as many languages as possible.

Mikolov et al. (2013b) first introduced a mapping strategy for aligning two monolingual spaces that learns a linear transformation from source to target space using stochastic gradient descent. This approach was later enhanced with the use of least squares for finding the optimal solution, L2-normalizing the word embedding, or constraining the mapping matrix to be orthogonal (Dinu et al., 2014; Shigeto et al., 2015; Xing et al., 2015; Artetxe et al., 2016; Smith et al., 2017); enhancements that soon became standard in the area. These models, however, are affected by **hubness**, where some words tend to be in the neighborhood of an exceptionally large number of other words, causing problems when using nearest-neighbor as the retrieval algorithm, and **neutrality**, where the resultant crosslingual space is highly conditioned by the characteristics of the language used as target. Hubness was addressed by a correction applied to nearest-neighbor retrieval whether using a inverted softmax (Smith et al., 2017) or a cross-domain similarity local scaling (Conneau et al., 2017) later incorporated as part of the training loss (Joulin et al., 2018). Neutrality was noticed by Doval et al. (2018), for which they proposed using two independent linear transformations so that the resulting crosslingual space is in a *middle* point between the two languages rather than just on the target language, and therefore not biased towards either language.

Other important trends in the area concentrate on (i) the search of unsupervised techniques for learning mapping functions (Conneau et al., 2017; Artetxe et al., 2018) and their versatility in dealing with low-resource languages (Vulić et al., 2019); (ii) the long-tail problem, where most existing crosslingual embedding generation strategies tend to under-perform (Braune et al., 2018; Czarnowska et al., 2019); and (iii) the formulation of more robust evaluation procedures oriented to determining the quality of generated crosslingual spaces (Glavas et al., 2019; Litschko et al., 2019).

Most existing works focus on a bilingual scenario. Yet, there is an increase on the interest for designing strategies that directly consider more than two languages at training time, thus creating fully multilingual spaces that do not depend on the

[1] Resources can be found at `https://github.com/ionmadrazo/HCEG`.

TB/MP model (Kementchedjhieva et al., 2018) for multilingual inference. Attempts to do so include the efforts by Søgaard et al. (2015), who leverage an inverted index based on the Wikipedia multilingual links to generate multilingual word representations. Wada et al. (2019) instead use a sentence-level neural language model for directly learning multilingual word embeddings and as a result bypassing the need for mapping functions. In the paradigm of aligning pre-trained word embeddings where we focus, Heyman et al. (2019) propose a technique that iteratively builds a multilingual space starting from a monolingual space and incrementally incorporating languages to it. Even if this strategy deviates from the traditional TB/MP model, it still preserves the idea of having a pivot language. Chen and Cardie (2018) separate the mapping functions into encoders and decoders, which are not language-pair dependent, unlike those in the TB/MP model. This removes the need for a pivot language, given that the multilingual space is now latent among all encoder and decoders and not centered in a specific language. The same pivot-removal effect is achieved by the strategy introduced in Jawanpuria et al. (2019), which generalizes a bilingual word embedding strategy into a multilingual counterpart by inducing a Mahalanobis similarity metric in the common space. These two strategies, however, still consider all languages equidistant to each other, ignoring the similarities and differences that lay among them.

Our work is inspired by Doval et al. (2018) and Chen and Cardie (2018), in the sense that it focuses on obtaining a non-biased or neutral crosslingual space that does not need to be centered in English (or any other pivot language) as the primary source. This neutrality is obtained by a compositional mapping strategy that hierarchically combines mapping functions in order to generate a single, non-language-centered crosslingual space, enabling a better mapping for languages that are distant or non-typologically related to English.

3 Proposed Strategy

A language family tree is a natural categorization of languages that has historically been used by linguistics as a reference that encodes similarities and differences across languages (Comrie, 1989). For example, based on the relative distances among languages in the tree illustrated in Figure 1, we infer that both Spanish and Portuguese are relatively similar to each other, given that they are part of the same Italic family. At the same time, both languages are farther apart from English than each other, and are radically different with respect to Finnish.

A language family tree offers a natural organization that can be exploited when building crosslingual spaces that integrate typologically diverse languages. We leverage this structure in HCEG, in order to generate a hierarchically compositional crosslingual word embedding space. Unlike traditional TB/MP strategies that generate a single crosslingual space, the result of HCEG is a set of transformation matrices that can be used to hierarchically compose the space required in each use-case. This maximizes the typological intra-similarity among languages used for generating the embedding space, while minimizing the differences across languages that can hinder the quality of the crosslingual embedding space. Thus, if an external application only considers languages that are Germanic, then it can just use the Germanic crosslingual space generated by HCEG, whereas if it needs languages beyond Germanic it can utilize a higher level family, such as the Indo-European. This cannot be done with the traditional TB/MP model. In this case, if an application is, for example, using only Uralic languages, then it would be forced to use an English-centered crosslingual space; this would in a decrease in the quality of the crosslingual space used because of the potential bad quality of mappings between typologically different languages, such as Uralic and Indo-European languages (Artetxe et al., 2017).

3.1 Definitions

Let $L = \{l_1, \ldots, l_{|L|}\}$ be a set of languages considered, $F = \{f_1, \ldots, f_{|F|}\}$ a set of language families, and $S = L \cup F = \{s_1, \ldots, s_{|F|+|L|}\}$ a set of possible language spaces. Let $X_l \in \mathbb{R}^{V_l \times d}$ be the set of word embeddings in language l, where V_l is the vocabulary of l and d is the number of dimensions of each embedding. Consider T as a language family tree (exemplified in Figure 1). The nodes in T represent language spaces in S, while each edge represents a transformation between the two nodes attached to it—that is, $W_{s_a \leftarrow s_b} \in \mathbb{R}^{d \times d}$ refers to the transformation from

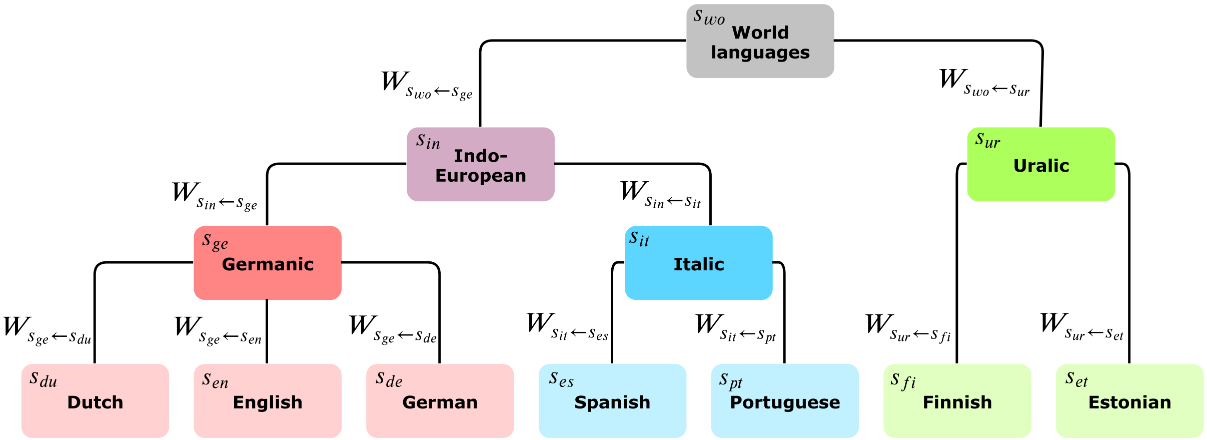

s_{wo} World languages
$W_{s_{wo}\leftarrow s_{ge}}$
$W_{s_{wo}\leftarrow s_{ur}}$
s_{in} Indo-European
s_{ur} Uralic
$W_{s_{in}\leftarrow s_{ge}}$
$W_{s_{in}\leftarrow s_{it}}$
s_{ge} Germanic
s_{it} Italic
$W_{s_{ge}\leftarrow s_{du}}$
$W_{s_{ge}\leftarrow s_{en}}$
$W_{s_{ge}\leftarrow s_{de}}$
$W_{s_{it}\leftarrow s_{es}}$
$W_{s_{it}\leftarrow s_{pt}}$
$W_{s_{ur}\leftarrow s_{fi}}$
$W_{s_{ur}\leftarrow s_{et}}$
s_{du} Dutch
s_{en} English
s_{de} German
s_{es} Spanish
s_{pt} Portuguese
s_{fi} Finnish
s_{et} Estonian

Figure 1: Sample language tree representation simplified for illustration purposes (Lewis and Gary, 2015).

space s_b to space s_a. For notation ease, we refer to $W_{s_a \xleftarrow{*} s_b}$ as the transformation that results from aggregating all transformations in the path from s_b to s_a, using the dot product:

$$W_{s_a \xleftarrow{*} s_b} = W_{s_a \leftarrow s_{t_1}} W_{s_{t_1} \leftarrow s_{t_2}} W_{s_{t_2} \leftarrow s_b} \quad (1)$$

where the path from s_a to s_b is $s_a, s_{t_1}, s_{t_2}, s_b$; s_{t_1} and s_{t_2} are intermediate spaces between s_a and s_b.

Finally, P is a set of bilingual lexicons, where $P_{l_1,l_2} \in \{0,1\}^{V_{l_1} \times V_{l_2}}$ is a bilingual lexicon with word pairs in languages l_1 and l_2. $P_{l_1,l_2}(i,j) = 1$ if the i^{th} word of V_{l_1} and the j^{th} word of V_{l_2} are aligned, $P_{l_1,l_2}(i,j) = 0$ otherwise.

Example. Consider the set of embeddings for English X_{en}, the transformation that converts embeddings in the English space to the Germanic language family space $W_{s_{ge} \xleftarrow{*} s_{en}}$, and the English embeddings transformed to the Germanic space $W_{s_{ge} \xleftarrow{*} s_{en}} X_{en}$. HCEG makes it so that $W_{s_{ge} \xleftarrow{*} s_{en}} X_{en}$ and $W_{s_{ge} \xleftarrow{*} s_{de}} X_{de}$ (the transformed embeddings of English and German) are in the same *Germanic* embedding space, while $W_{s_{in} \xleftarrow{*} s_{en}} X_{en}$ and $W_{s_{in} \xleftarrow{*} s_{es}} X_{es}$ (the transformed embeddings of English and Spanish) are in the same *Indo-European* embedding space.

In the rest of this section we describe HCEG in detail. Values given to each hyperparameter mentioned in this section are defined in Section 4.4.

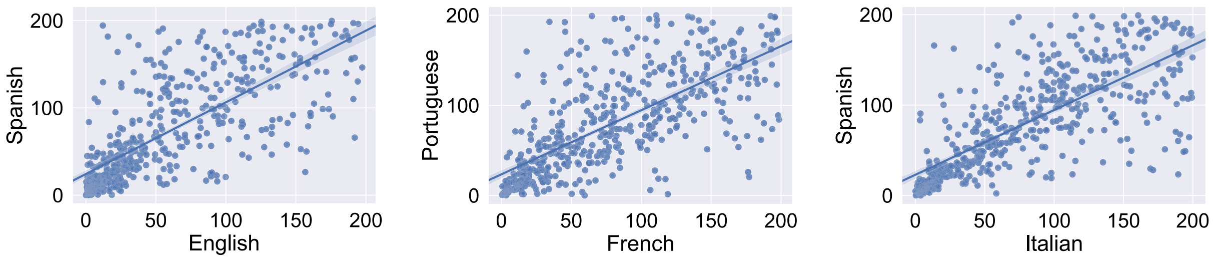
3.2 Embedding Normalization

When dealing with embeddings generated from different sources and languages, it is important to normalize them. For doing so, HCEG follows a normalization sequence shown to be beneficial (Artetxe et al., 2018), which consists of length normalization, mean centering, and a second length normalization. The last length normalization allows computing cosine similarity between embeddings in a more efficient manner, simplifying the computation of cosine similarity to a dot product given that the embeddings are of unit-length.

3.3 Word Pairs

In order to generate a crosslingual embedding space, HCEG requires a set P of aligned words across different languages. When using HCEG in a **supervised** way, P can be any existing resource consisting of bilingual lexicons, such as the ones described in Section 4.1. However, best advantage of the proposed strategy is taken when using **unsupervised** lexicon induction techniques, as they enable generating input lexicons for any pair of languages needed. Unlike TB/MP strategies that can only take advantage of signal that involves the pivot language, HCEG can use signal across all combinations of languages. For example, a TB/MP model where English is the pivot can only use lexicons composed of English words. Instead, HCEG can exploit bilingual lexicons from other languages, such as *Spanish-Portuguese* or *Spanish-Dutch*, that if using the language tree in Figure 1 would reinforce the training of $W_{s_{it} \leftarrow s_{es}}$, $W_{s_{it} \leftarrow s_{pt}}$ and $W_{s_{it} \leftarrow s_{es}}$, $W_{s_{in} \leftarrow s_{it}}$, $W_{s_{in} \leftarrow s_{ge}}$, $W_{s_{ge} \leftarrow s_{du}}$, respectively.

When using HCEG in unsupervised mode, P needs to be automatically inferred. Yet, computing each $P_{l_1,l_2} \in P$ given two monolingual embedding matrices X_{l_1} and X_{l_2} is not a trivial task, as X_{l_1} and



(a) $\rho = 0.684$ (b) $\rho = 0.709$ (c) $\rho = 0.737$

Figure 2: Distributions of word rankings across languages. The coordinates of each dot (representing a word pair) are determined by the position in the frequency ranking the word pair in each of the languages. Numbers are written in thousands. Scores computed using FastText embedding rankings (Grave et al., 2018) and MUSE crosslingual pairs (Conneau et al., 2017). Pearson's correlation (ρ) computed using the full set of word pairs, figures generated using a random sample of 500 word pairs for illustration purposes.

X_{l_2} are not aligned in vocabulary or dimension axes. Artetxe et al. (2018) leverage the fact that the relative distances among words are maintained across languages (Mikolov et al., 2013b), and thus propose using a language-agnostic representation M_l for generating an initial alignment P_{l_1,l_2}:

$$M_l = sorted(X_l X_l^\top) \qquad (2)$$

where given that X_l is length normalized, and $X_l X_l^\top$ computes a matrix of dimensions $V_l \times V_l$ containing in each row the cosine similarities of the corresponding word embedding with respect to all other word embeddings. The values in each row are then *sorted* to generate a distribution representation of each word that in a ideal case where the isometry assumption holds perfectly would be language agnostic. Using the embedding representations M_{l_1} and M_{l_2}, P_{l_1,l_2} can be computed by assigning each word its most similar representation as its pair, that is, $P_{l_1,l_2}(i,j) = 1$ if:

$$j = \arg \max_{1 \le j \le V_l} M_{l_1}(i,*) M_{l_2}(j,*)^\top \qquad (3)$$

where $M_{l_1}(i,*)$ is the i^{th} row of M_{l_1} and $M_{l_2}(j,*)$ is the j^{th} row of M_{l_2}.

The results in Artetxe et al. (2018) indicate that this assumption is strong enough to generate an initial alignment across languages. However, as we demonstrate in Section 3.3, the quality of this type of initial alignment is dependent on the languages used, making this initialization not applicable for languages that are typologically too distant from each other—a statement also echoed by Artetxe et al. (2018) and Søgaard et al. (2018).

To ensure a more robust initialization, we enhance the strategy presented in Artetxe et al. (2018) by introducing a new signal based on the frequency of use of words. Lin et al. (2012) found that the top-2 most frequent words tend to be consistent across different languages. Motivated by this result, we measure to what extent the frequency rankings of words correlates across languages. As shown in Figure 2, the word-frequency rankings are strongly correlated across languages, meaning that popular words tend to be popular regardless of the language. We exploit this behavior in order to reduce the search space of Equation (3) as follows:

$$j = \arg \max_{j-t \le j \le j+t} M_{l_1}(i,*) M_{l_2}(j,*)^\top \qquad (4)$$

where t is a value used to determine the search window. Note that we assume the embeddings in any matrix X_l are sorted in ascending order of frequency, namely, the embedding in the first row represents the most frequent word of language l. Apart from improving the overall quality of the inferred lexicons (see Section 5.1), incorporating a frequency ranking based search as part of the initialization reduces the computation time needed as the search space is considerably reduced.

3.4 Objective Function

Unlike traditional objective functions that optimize a transformation matrix for two languages at a time, the goal of HCEG is to simultaneously optimize the set of all transformation matrices W such that the loss function $\mathcal{L}$ is minimized:

$$\arg \min_{W} \mathcal{L} \qquad (5)$$

$\mathcal{L}$ is a linear combination of three different losses:

$$\mathcal{L} = \beta_1 \times \mathcal{L}_{align} + \beta_2 \times \mathcal{L}_{orth} + \beta_3 \times \mathcal{L}_{reg} \quad (6)$$

where $\mathcal{L}_{align}$, $\mathcal{L}_{orth}$, $\mathcal{L}_{reg}$, represent the alignment, orthogonality, and regularization losses, and β_1, β_2, β_3 are their weights.

$\mathcal{L}_{align}$ gauges the extent to which training word pairs align. This is done by computing the sum of the cosine similarity among all word pairs in P:

$$\begin{aligned}\mathcal{L}_{align} = -\sum_{P_{l_1,l_2} \in P} P_{l_1,l_2}(W_{s_{\widehat{l_1,l_2}} \overset{*}{\leftarrow} s_{l1}} X_{l_1} \cdot \\ W_{s_{\widehat{l_1,l_2}} \overset{*}{\leftarrow} s_{l2}} X_{l_2})\end{aligned} \quad (7)$$

where $s_{\widehat{l_1,l_2}}$ refers to the space in the lowest common parent node for s_{l_1} and s_{l_2} in T (e.g., $s_{\widehat{es,en}} = s_{in}$ in Figure 1). We found that using $s_{\widehat{l_1,l_2}}$ instead of the space in the root node of T improves the overall performance of HCEG, apart from reducing the time taken for training (see Section 5.3).

Several researchers have found it beneficial to enforce orthogonality in the transformation matrices W (Xing et al., 2015; Artetxe et al., 2016; Smith et al., 2017). This constraint ensures that the original quality of the embeddings is not degraded when transforming them to a crosslingual space. For this reason, we incorporate an orthogonality constraint $\mathcal{L}_{orth}$ into our loss function in Equation 8, with I being the identity matrix.

$$\mathcal{L}_{orth} = \sum_{W_{s_1 \leftarrow s_2} \in W} \| I - W_{s_1 \leftarrow s_2} W_{s_1 \leftarrow s_2}^{\top} \| \quad (8)$$

We also find it beneficial to include a regularization term in $\mathcal{L}$:

$$\mathcal{L}_{reg} = \sum_{W_{s_1 \leftarrow s_2} \in W} \| W_{s_1 \leftarrow s_2} \|_2 \quad (9)$$

3.5 Learning the Parameters

HCEG utilizes stochastic gradient descent for tuning the parameters in W with respect to the training word pairs in P. In each iteration, $\mathcal{L}$ is computed and backtracked in order to tune each transformation matrix in W such that $\mathcal{L}$ is minimized. Batching is used to reduce the computational load in each iteration. A batch of word pairs $\hat{P}$ is sampled from P by randomly selecting α_{lpairs} language pairs as well as α_{wpairs} word pairs in each $\hat{P}_{l_1,l_2} \in \hat{P}$—for example, a batch might consist of 10 $\hat{P}_{l_1,l_2}$ matrices each containing 500 aligned words.

Iterations are grouped into epochs of α_{iter} iterations at the end of which $\mathcal{L}$ is computed for the whole P. We take a conservative approach as convergence criterion. If no improvement is found in $\mathcal{L}$ in the last α_{conv} epochs, the training loop stops.

We achieve best convergence time initializing each $W_{s_1 \leftarrow s_2} \in W$ to be orthogonal. We tried several methods for orthogonal initialization, such as simply initializing to the identity matrix. However, we obtained most consistent results using the random semi-orthogonal initialization introduced by Saxe et al. (2013).

3.6 Iterative Refinement

As shown by Artetxe et al. (2017), the initial lexicon P is iteratively improved by using the generated crosslingual space for inferring a new lexicon P' at the end of each learning phase described in Section 3.5. More specifically, when computing each $P'_{l_1,l_2} \in P'$, $P'_{l_1,l_2}(i,j)$ is 1 (0 otherwise) if

$$\begin{aligned} j = \arg\max_{j} \; & W_{s_{\widehat{l_1,l_2}} \overset{*}{\leftarrow} s_{l1}} X_{l_1}(i,*) \cdot \\ & (W_{s_{\widehat{l_1,l_2}} \overset{*}{\leftarrow} s_{l2}} X_{l_2}(j,*))^{\top}\end{aligned} \quad (10)$$

Potentially, any new bilingual lexicon P'_{l_1,l_2} can be inferred and included in P' at the end of each learning phase. However, as the cardinality of L grows, this process can take a prohibitive amount of time given combinatorial explosion. Therefore, in practice, we only infer P'_{l_1,l_2} following a criterion intended to maximize lexicon quality. P'_{l_1,l_2} is inferred for languages l_1 and l_2 only if l_1 and l_2 are siblings in T (they share the same parent node) or l_1 and l_2 are the best representatives of their corresponding family. A language is deemed the *best representative* of its family if it is the most frequently-spoken[2] language in its subtree. For example, in Figure 1, Spanish is the *best* representative for the Italic family, but not for Indo-European, for which English is used.

The set criterion not only reduces the amount of time required to infer P' but also improves overall HCEG performance. This is due to a better utilization of the hierarchical characteristics of our crosslingual space, only inferring bilingual lexicons from typologically related languages or

[2] Based on numbers reported by Lewis and Gary (2015).

their best representatives in terms of resource quality.

3.7 Retrieval Criterion

As discussed in Section 2, one of the issues effecting nearest-neighbor retrieval is hubness (Dinu et al., 2014), where certain words are in the surrounding of an abnormally large number of other words, causing the nearest-neighbor algorithm to incorrectly prioritize hub words. To address this issue, we use Cross-domain Similarity Local Scaling (CSLS) (Conneau et al., 2017) as the retrieval algorithm during both training and prediction time. CSLS is a rectification for nearest-neighbor retrieval that avoids hubness by counterbalancing the cosine similarity between two embeddings by a factor consisting of the average similarity of each embeddings with its k closest neighbors. Following the criteria in Conneau et al. (2017), we set the number of neighbours used by CSLS to $k = 10$.

4 Evaluation Framework

We describe below the evaluation set up used for conducting the experiments presented in Section 5.

4.1 Word Pair Datasets

Dinu-Artetxe. The Dinu-Artetxe dataset, presented by Dinu et al. (2014) and enhanced by Artetxe et al. (2016), is the one of the first benchmarks for evaluating crosslingual embeddings. It is composed of English-centered bilingual lexicons for Italian, Spanish, German, and Finnish.

MUSE. The MUSE dataset (Conneau et al., 2017) contains bilingual lexicons for all combinations of German, English, Spanish, French, Italian, and Portuguese. In addition, it includes word pairs for 44 languages with respect to English.

Panlex. Dinu-Artetxe and MUSE are both English-centered datasets, given that most (if not all) of their word pairs have English as their source or target language. This makes the datasets suboptimal for our purpose of generating and evaluating a non-language centered crosslingual space. For this reason, we generated a dataset using Panlex (Kamholz et al., 2014), a panlingual lexical database. This dataset (made public in our repository) includes bilingual lexicons for all combinations of 157 languages for which FastText is available, totalling 24,492 bilingual lexicons. Each of the lexicons was generated by randomly sampling 5k words from the top-200k words in the embedding set for the source language, and translating them to the target language using the Panlex database. We find it important to highlight that this dataset contains considerably more noise than other datasets given that Panlex is generated in an automatic way and is not as finely curated by humans as previous datasets. We still find comparisons using this dataset fair, given that its noisy nature should affect all strategies equally.

4.2 Language Selection and Family Tree

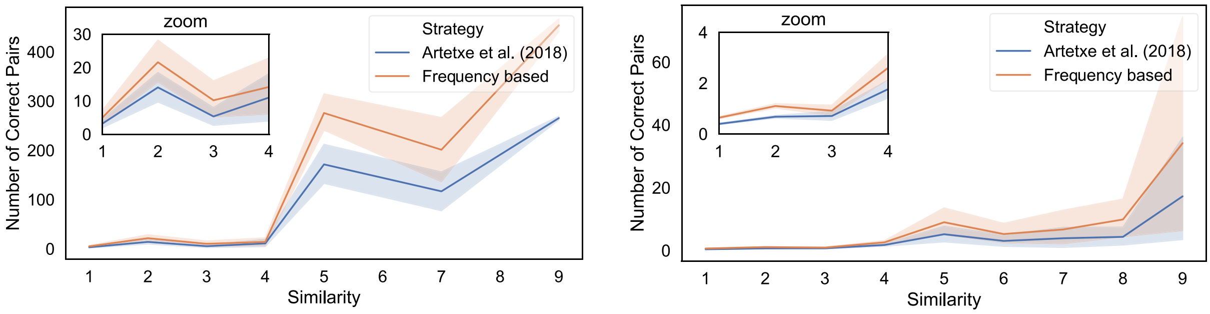
As previously stated, we aim to generate a single crosslingual space for as many languages as possible. We started with the 157 languages for which FastText embeddings are available (Grave et al., 2018). We then removed languages that did not meet both of the following criteria: (1) there must exist a bilingual lexicon with at least 500 word pairs for the language in any of the datasets described in Section 4.1, and (2) the embedding set provided by FastText must contain at least 20k words. The first criterion is a minimal condition for evaluation, while the second one is necessary for the unsupervised initialization strategy. The criteria are met by 107 languages, which are the ones used in our experiments. Their corresponding ISO-639 codes can be seen later in Table 5. We use the language family tree defined by Lewis and Gary (2015).

4.3 Framework

For experimental purposes, each dataset described in Section 4.1 is split into training and testing sets. We use the original train-test splits for Dinu-Artetxe and MUSE. For Panlex, we generate a split randomly sampling word pairs—keeping 80% for the training and the remaining 20% for testing. For development and parameter tuning purposes, we use a disjoint set of word pairs specifically created for this purpose based on the Panlex lexical database. This development set contains 10 different languages with varied popularity. None of the word pairs present in this development set are part of either the train or test sets.

4.4 Hyperparameters

The following hyperparameters were manually tuned using the development set described in Section 4.3: $\beta_1 = 0.98$, $\beta_2 = 0.01$, $\beta_3 = 0.01$,



(a) MUSE dataset

(b) Panlex dataset

Figure 3: Number of correct word pairs inferred using the unsupervised initialization technique presented by Artetxe et al. (2018) and the Frequency based technique described in Section 3.3.

$t = 1000$, $\alpha_{lpairs} = 128$, $\alpha_{wpairs} = 2048$, $\alpha_{iter} = 5000$, $\alpha_{conv} = 25$.

5 Evaluation

We discuss below the results of the study conducted over 107 languages to assess HCEG.

5.1 Unsupervised Initialization

We first evaluate the performance of the unsupervised initialization strategy described in Section 3.3, and compare it with the state-of-the-art strategy proposed by Artetxe et al. (2018). In this case, we run both initialization strategies using the top-20k FastText embeddings (Grave et al., 2018) for all pairwise combinations of the 107 languages we study. For each language pair, we measure how many of the inferred word pairs are present in the corresponding lexicons in the MUSE and Panlex datasets. For MUSE, our proposed initialization strategy (*Frequency based*) obtains an average of 48.09 correct pairs, an improvement with respect to the 29.62 obtained by the strategy proposed by Artetxe et al. (2018). For Panlex, the respective average correct pair counts are 1.05 and 0.55. Both differences are statistically significant ($p < 0.01$) using a paired t-test. The noticeable difference across datasets is due to how the sampling was done for generating the datasets: MUSE contains a considerably higher number of frequent words in comparison to Panlex, making the latter a relatively harder dataset for vocabulary induction. In Figure 3 we illustrate the results of each strategy grouped by language-pair similarity. This similarity is based on the number of common parents the two languages share. For example, in Figure 1, Spanish has a similarity of 3, 2, and 1 with Portuguese, English, and Finnish, respectively. As we see in Figure 3, similarity is a factor that strongly determines the quality of the alignment generated by the unsupervised initialization. Even if this phenomenon affects both analyzed strategies, our proposed frequency-based initialization strategy consistently obtains a few more correct word pairs for the least similar language pairs, which, as we show in Table 4, are key for generating a correct mapping for those languages.

5.2 State-of-the-Art Comparison

In order to contextualize the performance of HCEG with respect to the state-of-the-art (listed in Tables 1 and 2), we measure the word translation prediction capabilities of each of the strategies. We do so using **Precision@1** for bilingual lexicon induction as a means to quantify vocabulary induction performance. Scores reported hereafter are average Precision@1 in percentage form, for each of the words in the testing set.

When applicable, we report results for both the supervised (HCEG-S) and unsupervised (HCEG-U) versions of HCEG. In the supervised mode, we train one single model per dataset using all the training word pairs available. We then use this model for computing all pairwise scores. In the unsupervised mode, unless explicitly stated otherwise, we train a single model regardless of the dataset used for testing purposes. This means that, in some cases, the unsupervised mode leverages monolingual data beyond the languages used for testing, as it uses all 107 language embeddings. We found it unfair to train a supervised model using

	Method	en-it	en-de	en-fi	en-es
Supervised	Mikolov (2013b)	34.93*	35.00*	25.91*	27.73*
	Faruqui (2014)	38.40*	37.13*	27.60*	26.80*
	Shigeto (2015)	41.53*	43.07*	31.04*	33.73*
	Dinu (2014)	37.7	38.93*	29.14*	30.40*
	Lazaridou (2015)	40.2	-	-	-
	Xing (2015)	36.87*	41.27*	28.23*	31.20*
	Zhang (2016)	36.73*	40.80*	28.16*	31.07*
	Artetxe (2016)	39.27	41.87*	30.62*	31.40*
	Artetxe (2017)	39.67	40.87	28.72	-
	Smith (2017)	43.1	43.33*	29.42*	35.13*
	Artetxe (2018a)	45.27	44.13	32.94	36.60
	Jouling (2018)	45.5	-	-	-
	Jawanpuria (2019) mul	48.7	49.1	36.0	36.0
	Jawanpuria (2019)	48.3	**49.3**	**36.1**	39.3
Semi.	Artetxe (2017) 25	37.27	39.60	28.16	-
	Smith (2017) cog	39.9	-	-	-
	Artetxe (2017) num	39.40	40.27	26.47	-
Unsupervised	Zhang (2017), $\lambda = 1$	0.00*	0.00*	0.00*	0.00*
	Zhang (2017) $\lambda = 10$	0.00*	0.00*	0.01*	0.01*
	Conneau (2017) code	45.15*	46.83*	0.38*	35.38*
	Conneau (2017) paper	45.1	0.01*	0.01*	35.44*
	Artetxe (2018)	48.13	48.19	32.63	37.33
	HCEG-U	**49.02**	48.18	34.82	**42.15**

Table 1: Results using the Dinu-Artetxe dataset. Scores marked with (*) were reported by Artetxe et al. (2018); the remaining ones were reported in the corresponding original papers.

the Dinu-Artetxe dataset given that it only contains four bilingual lexicons, not enough for training our tree structure. Thus, only unsupervised results are shown for that dataset.

As shown in Table 1, the unsupervised version of HCEG achieves, in most cases, the best performance among all unsupervised strategies, even improving over state-of-the-art supervised models in some cases. The improvement is most noticeable for Italian and Spanish, where HCEG-U obtains an improvement of 1 and 3 points, respectively. A similar behavior can be seen in Table 2, where we describe the results on the MUSE dataset. Spanish, along with Catalan, Italian, and Portuguese, obtains a substantially larger improvement compared with other languages. We attribute this to the fact that Spanish is the second most resourceful language in terms of corpora after English. This makes the quality of Spanish word embeddings comparably better than other languages, which as a result improves the mapping quality of typologically related languages, such as Portuguese, Italian, or Catalan.

To further contextualize the performance of HCEG-U, in terms of its capability for generating crosslingual embeddings in an unsupervised fashion, we conducted further experiments. In Table 3, we summarize the results obtained from comparing HCEG-U with other unsupervised strategies focused on learning crosslingual word embeddings. In our comparisons we include (i) a direct bilingual learning baseline that simply learns a bilingual mapping using two monolingual word embeddings (Conneau et al., 2017), (ii) a pivot-based strategy that can leverage a third language for learning a crosslingual space (Conneau et al., 2017), and (iii) a fully multilingual, pivotless strategy that aggregates languages into a joint space in an iterative manner (Chen and

	Conneau (2017)	Joulin (2018)	Artetxe (2018)	HCEG-S	HCEG-U^{-}	HCEG-U
bg	57.5	63.9	65.8	64.1	64.0	**67.5**
ca	70.9	73.8	76.3	73.1	74.2	**77.7**
cs	64.5	68.2	70.2	68.2	65.9	**71.7**
da	67.4	71.1	70.3	68.8	71.9	**72.7**
de	72.7	76.9	**79.1**	75.8	75.2	79.0
el	58.5	62.7	67.8	65.3	66.4	**68.5**
es	83.5	86.4	88.6	86.8	86.4	**90.4**
et	45.7	49.5	55.8	53.5	53.4	**57.3**
fi	59.5	65.8	68.1	65.2	65.4	**68.3**
fr	82.4	84.7	87.6	85.4	85.7	**88.3**
he	54.1	57.8	61.1	59.5	61.4	**63.0**
hr	52.2	55.6	57.6	54.8	54.1	**58.2**
hu	64.9	69.3	69.6	66.8	64.5	**70.1**
id	67.9	69.7	75.5	73.2	73.5	**75.6**
it	77.9	81.5	83.3	81.3	79.7	**85.6**
mk	54.6	59.9	63.5	62.3	62.5	**64.9**
nl	75.3	79.7	79.9	79.4	79.7	**81.9**
no	67.4	71.2	69.9	69.5	69.3	**71.9**
pl	66.9	70.5	72.0	70.7	70.5	**72.8**
pt	80.3	82.9	85.5	83.8	83.2	**87.8**
ro	68.1	74.0	75.4	72.8	71.7	**76.0**
ru	63.7	67.1	69.5	68.1	69.1	**69.8**
sk	55.3	59.0	62.0	59.6	56.7	**62.4**
sl	50.4	54.2	60.1	57.7	59.7	**61.1**
sv	60.0	63.7	66.2	65.0	64.8	**68.0**
tr	59.2	61.9	68.7	66.3	66.6	**70.0**
uk	49.3	51.5	**56.4**	53.8	55.7	**56.4**
vi	55.8	55.8	3.9	55.5	55.6	**58.3**
Avg.	63.8	67.4	68.2	68.1	68.1	**71.2**

Table 2: Results on the MUSE dataset. Scores from Artetxe et al. (2018) were obtained using the scripts shared by the authors. All the other scores were reported in Joulin et al. (2018). HCEG-U^{-} only considers the 29 languages in the experiment for training.

Cardie, 2018). From the reported results, we see that HCEG-U^{-} outperforms all other considered strategies for 24 out of 30 language pairs. Highest improvements are found for languages of the Italic family (Spanish, Portuguese, Italian, and French). We observe that HCEG-U^{-} under-performed when the corresponding experiment involved the German language as source or target. We attribute this behavior to the fact that the Italic family is predominant in the languages explored in this experiment.

In order to perform a fair comparison with respect to the work proposed by Chen and Cardie (2018), we limited the monolingual data that HCEG-U^{-} used to the six languages considered in this experiment (results that are reported in Table 3). However, in order to show the full potential of HCEG-U, we also include results achieved when using 107 languages (column HCEG-U). As seen in Tables 2 and 3, the differences between HCEG-U^{-} and HCEG-U are considerable, manifesting the capabilities of the proposed model to take advantage of monolingual data in multiple languages at the same time.

The importance of explicitly considering topological connections among languages to

Method	Type	en-de	en-fr	en-es	en-it	en-pt	de-fr	de-es	de-it	de-pt	fr-es	fr-it	fr-pt	es-it	es-pt	it-pt
Conneau (2017)	Direct	74.0	82.3	81.7	77.0	80.7	73.0	65.7	66.5	58.5	83.1	83.0	77.9	83.3	87.3	80.5
Conneau (2017)	Pivot	74.0	82.3	81.7	77.0	80.7	71.9	66.1	68.0	57.4	81.1	79.7	74.7	81.9	85.0	78.9
Chen (2018)	Multi	**74.8**	82.4	82.5	78.8	81.5	**76.7**	**69.6**	72.0	63.2	83.9	83.5	79.3	84.5	87.8	82.3
HCEG-U$^-$	Multi	74.5	**82.8**	**82.7**	**79.5**	**81.7**	73.5	68.0	**72.2**	**63.3**	**84.4**	**83.9**	**79.8**	**86.0**	**88.9**	**83.6**
HCEG-U	Multi	79.4	88.4	89.8	85.4	88.1	77.4	72.3	76.5	66.7	89.1	86.1	84.8	89.4	89.7	86.3

Method	Type	de-en	fr-en	es-en	it-en	pt-en	fr-de	es-de	it-de	pt-de	es-fr	it-fr	pt-fr	it-es	pt-es	pt-it
Conneau (2017)	Direct	72.2	82.1	83.3	77.7	80.1	69.7	68.8	62.5	60.5	86	87.6	83.9	87.7	92.1	80.6
Conneau (2017)	Pivot	72.2	82.1	83.3	77.7	80.1	68.1	67.9	66.1	63.1	84.7	86.5	82.6	85.8	91.3	79.2
Chen (2018)	Multi	**72.9**	81.8	83.7	77.4	79.9	**71.2**	**69.0**	69.5	**65.7**	86.9	88.1	86.3	88.2	92.7	82.6
HCEG-U$^-$	Multi	72.4	**82.6**	**84.1**	**77.8**	**80.3**	**71.2**	67.8	**69.6**	65.6	**87.5**	**88.8**	**87.0**	**89.5**	**94.0**	**83.9**
HCEG-U	Multi	78.6	88.2	91.0	85.8	87.5	75.4	71.2	73.9	68.6	90.6	91.0	90.2	91.4	94.3	87.1

Table 3: Comparison of unsupervised crosslingual embedding learning strategies under different merging scenarios in the MUSE dataset. *Direct* indicates a traditional bilingual scenario where a mapping from source to target is learned. *Pivot* uses an auxiliary pivot language (English) for merging multiple languages into the same space. *Multi* merges all languages into the same space without using a pivot. All scores except HCEG-U were originally reported by Chen and Cardie (2018). HCEG-U$^-$ only considers the six languages in the experiment for training. Note that HCEG-U is excluded when highlighting the best model (**bold**), given that it uses monolingual data beyond what other models do.

enhance mappings become more evident when analyzing the data in Table 5. Here we include the pairing that yielded the best and worst mapping for each language, as well as the position of English in the quality ranking. English and Spanish have a strong quality mapping with respect to each other, Spanish being the language with which English obtains the best mapping and English is the second-best mapped language for Spanish. Additionally, Spanish is the language with which Italian, Portuguese, and Catalan obtain the best mapping quality. On the other side of the spectrum, the worst mappings are dominated by two languages, Georgian and Vietnamese, with 40 languages having these two language as worst; this is followed by Maltese, Albanian, and Finnish, with 8 occurrences each. This is not unexpected, as these languages are relatively isolated in the language family tree, and also have a low number of speakers. We also see that English is usually on the top side of the ranking for most languages. For languages that are completely isolated, such as Basque and Yoruba, English tends to be their best mapped language. From this we surmise that when typological relations are lacking, the quality of the embedding space is the only aspect the mapping strategy can rely on.

Given space constraints, we cannot show the vocabulary induction scores for the 24,492 language pairs in the Panlex dataset. Instead, we group the results using two variables: the sum of number of speakers for each of the two languages, and the minimum similarity (as defined in Section 5.1) for each language with respect to English. We rely on these variables for grouping purposes as they align with two of our objectives for designing HCEG: (1) remove the bias towards the pivot language (English), and (2) improve the performance of low-resource languages by taking advantage of typologically similar languages.

Figure 4 captures the improvement (2.7 on average) of HCEG-U over the strategy introduced in Artetxe et al. (2018) (the best-performing benchmark), grouped by the aforementioned variables. We excluded Hindi and Chinese from the figure, as they made any pattern hard to observe given their high number of speakers. The sum of number of speakers axis was also logarithmically scaled to facilitate visualization. The figure captures an evident trend in the similarity axis. The lower the similarity of the language with respect to English, the higher the improvement achieved by HCEG-U. This can be attributed to the manner in which TB/MP models generated the space using English as primary resource, hindering the potential quality of languages that are distant from it. Additionally, we see a less-prominent but existing trend in

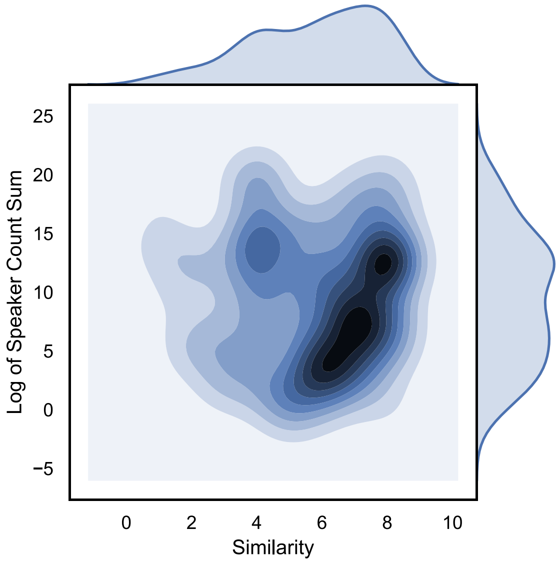


Figure 4: Improvement over the strategy proposed by Artetxe et al. (2018) in Panlex, in terms of language similarity and number of speakers. Darker denotes larger improvement.

	Description	Dinu-Artetxe	MUSE	Panlex
Supervised	¬Hierarchy	-	66.7	32.0
	¬Orthogonal Init.	-	67.8	36.5
	¬Iterative Refinement	-	65.4	35.1
	All vs All Inference	-	66.3	36.6
	World langs. as root	-	67.5	35.7
	HCEG-S	-	**68.1**	**37.3**
Unsupervised	¬Hierarchy	40.2	67.9	28.1
	¬Orthogonal Init.	43.2	71.0	34.7
	¬Iterative Refinement	0.09	0.08	0.02
	All vs All Inference	39.3	69.4	34.6
	World langs. as root	42.8	70.2	33.8
	¬Freq. based Init.	41.2	68.0	31.1
	HCEG-U	**43.5**	**71.2**	**35.8**

Table 4: Ablation study.

the speaker sum axis. Despite some exceptions, HCEG-U obtains higher differences with respect to Artetxe et al. (2018) the less spoken a language is. A behavior that is similar in essence to a Pareto front can also be depicted from the figure. Even if both variables contribute to the difference in improvement of HCEG-U, one variable needs to compensate for the other in order to maximize accuracy. In other words, the improvement is higher the fewer speakers the language pair has or the more distant the two languages are from English, but when both variables go to the extreme, the improvement decreases. The aforementioned trends serve as evidence that the hierarchical structure is indeed important when building a crosslingual space that considers typologically diverse languages, validating our premises for designing HCEG.

5.3 Ablation Study

In order to assess the validity of each functionality included as part of HCEG, we conducted an ablation study. We summarize the results of this study in Table 4, where the symbol ¬ indicates that the subsequent feature is ablated in the model. For example, ¬Hierarchy indicates that the Hierarchy structure is removed, replacing it by a structure where each language needs just one transformation matrix to reach the *World languages* space.

As indicated by the ablation results, the hierarchical structure is indeed a key part of HCEG, considerably reducing its performance when removed, and having its strongest effect in the dataset with the highest number of languages (i.e., Panlex). The importance of the Iterative Refinement strategy is also noticeable, making the unsupervised version of HCEG useless when removed. The Frequency-based initialization is also a characteristic that considerably improves the results of HCEG-U. Looking deeper into the data, we found 2,198 language pairs (about 9% of all pairs) that obtained a vocabulary induction accuracy close to 0 (<0.05) without using this initialization, but were able to produce enough signal to yield more substantial accuracy values (>10.0) when using the Frequency-based initialization. Finally, the design decisions that we initially took for reducing training time—(i) the orthogonal initialization, (ii) the heuristic based inference, and (iii) using the lowest common root for computing the loss function—also have a positive effect on the performance of the HCEG.

5.4 Influence of Pivot Choice

One of the premises for building HCEG was to design a strategy that would not require pivots for achieving a single space with multiple word embeddings, given that a pivot induces a bias into the final space that can hinder the quality of the mapping for languages that are too distant to it. In this section we describe the results of experiments conducted for measuring the effect pivot selection can have on the performance of the mapping. For doing so, we measure the

L	B,W,E	L	B,W,E	L	B,W,E	L	B,W,E	L	B,W,E	L	B,W,E	L	B,W,E
af	nl,fi,4	ceb	tl,li,22	ga	gd,tt,12	jv	id,scn,34	my	zh,mk,19	sco	en,mt,1	tr	tk,ka,13
als	en,vi,1	ckb	tg,tr,19	gd	ga,vi,2	ka	en,bs,1	nds	nl,vi,3	sd	bn,tl,5	tt	ba,sa,9
am	arz,de,80	cs	sk,vi,12	gl	pt,ka,16	kk	ky,vi,51	nl	af,ka,4	si	dv,ka,5	ug	tr,vls,4
an	es,ka,17	cv	tr,sq,2	gom	mr,fi,10	km	vi,nl,4	no	sv,vi,3	sk	cs,vi,5	uk	ru,fi,19
arz	mt,ja,3	cy	br,fi,2	gu	pa,ka,3	kn	ta,lt,55	oc	es,my,3	sl	sr,vi,6	ur	hi,eo,10
as	bn,vi,4	da	sv,fi,4	he	arz,mk,10	ko	en,af,1	pa	gu,vi,6	so	arz,sq,73	vec	pms,tr,2
ast	es,ja,20	de	lb,mt,5	hi	ur,ka,5	ky	kk,af,17	pam	id,sr,18	sq	en,tt,1	vi	km,vls,3
ba	tt,sq,34	dv	si,ka,3	hr	sr,tt,5	la	es,mt,3	pl	cs,vi,4	sr	hr,vi,4	vls	nl,eo,8
bar	de,fi,6	el	en,eo,1	hsb	pl,am,3	lb	de,ka,2	pms	vec,sah,7	su	id,mk,37	wa	fr,fi,7
be	ru,vi,4	en	es,gv,-	hu	fi,ckb,9	li	nl,ka,7	pt	es,mt,5	sv	da,vi,5	yo	en,lt,1
bg	mk,ka,9	eo	en,sq,1	hy	en,fi,1	lt	ru,mt,5	qu	en,bn,1	ta	ml,mt,3	zh	my,de,10
bn	as,vi,6	es	pt,vi,2	id	jv,vi,3	mg	id,sq,44	ro	es,vi,6	te	ta,mk,15		
br	cy,ka,18	eu	en,lt,1	ilo	id,sq,6	mk	bg,vi,4	ru	uk,su,20	tg	ckb,ka,13		
bs	sr,ka,2	fi	hu,als,24	is	sv,ka,3	ml	ta,sq,29	sa	hi,ka,2	th	en,vls,1		
ca	es,mt,5	fr	it,vi,5	it	es,mt,5	mr	si,ka,21	sah	tr,ka,2	tk	tr,lt,7		
ce	en,sq,1	fy	en,eo,1	ja	en,vi,1	mt	arz,tt,70	scn	it,ka,21	tl	ceb,ru,47		

Table 5: Best (B), worst (W), and English mapping ranking (E) for each language (L).

Pivot	Language Family	Afro-Asiatic	Austronesian	Indo-European/Balto-Slavic	Indo-European/Germanic	Indo-European/Indo-Iranian	Indo-European/Italic	Sino-tibetan	Turkic	Uralic	Avg.
en	Indo-European/Germanic	27.3	28.7	32.1	**39.8**	31.4	40.4	27.3	26.9	28.3	**31.4**
arz	Afro-Asiatic	**30.2**	27.1	28.1	32.1	28.3	33.4	25.1	23.4	27.1	28.3
id	Austronesian	27.1	**30.3**	27.7	31.1	28.3	32.5	25.8	24.6	27.6	28.3
ru	Indo-European/Balto-Slavic	26.3	26.3	**34.2**	38.2	28.5	37.3	24.6	22.5	26.8	29.4
de	Indo-European/Germanic	25.1	26.9	25.1	37.6	27.3	37.2	24.7	23.7	25.6	28.1
hi	Indo-European/Indo-Iranian	26.3	27.1	26.1	33.7	**32.3**	34.2	23.4	25.6	26.4	28.3
es	Indo-European/Italic	26.9	26.7	30.6	38.5	31.0	**41.5**	26.8	26.7	28.4	30.8
pt	Indo-European/Italic	26.0	26.6	30.4	37.9	27.7	41.3	25.9	26.4	26.5	29.9
zh	Sino-Tibetan	25.1	27.3	25.3	23.4	26.1	24.8	**29.3**	25.7	27.6	26.1
tr	Turkic	24.9	25.3	25.5	28.2	27.8	28.6	25.3	**28.7**	27.3	26.8
hu	Uralic	25.4	25.8	25.8	31.8	26.4	32.8	25.5	21.9	**30.1**	27.3

Table 6: Results obtained by existing bilingual mapping strategies using different pivots on the Panlex dataset. Values in each cell indicate the average performance obtained for each of the pairwise combinations of languages under the family noted in the corresponding column title. For example, the first cell indicates the average score obtained for all possible combinations of afro-asiatic languages using English as a pivot. Results are averaged across the strategy presented in Conneau et al. (2017) and Artetxe et al. (2018) in order to avoid system-specific biases.

performance of state-of-the-art bilingual mapping strategies in a pivot-based inference scenario. We use 11 different pivots and average the results of two different strategies—(Conneau et al., 2017) and (Artetxe et al., 2018)—grouped by several language families. As depicted by the results presented in Table 6, selecting a pivot that belongs to the family of the languages being tested is always the best choice. In cases where we considered multiple pivots of the same family, the most resource-rich language resulted in the best option, namely, Spanish in the case of the Italic family and English for the Germanic family. On average, English is the best choice of pivot if all language families need to be considered, followed by Spanish and Portuguese. This validates two of the design decisions for HCEG, that is, the need to avoid selecting a pivot and the importance of using the languages with largest speaker-base when performing language transfer.

6 Conclusion and Future Work

We have introduced HCEG, a crosslingual space learning strategy that does not depend on a pivot language, as instead, it takes advantage of the natural hierarchy existing among languages. Results from extensive studies on 107 languages demonstrate that the proposed strategy outperforms existing crosslingual space generation techniques, in terms of vocabulary induction, for both popular and not so popular languages. HCEG improves the mapping quality of many low-resource languages. We noticed that this improvement mostly happens when a language has more typologically related counterparts, however. Therefore, as future work, we intend to investigate other techniques that can help improve the quality of mapping for typologically isolated low-resource languages. Additionally, it is important to note that the time complexity required by the proposed algorithm is $N(N-1)$, with N being the number of languages considered. For the traditional TB/MP strategy, complexity is limited to learning from N language pairs. Therefore, we plan on exploring strategies to reduce the number of language pairs that need to be learned for creating the crosslingual space. Finally, we will explore different data-driven strategies for building the tree structure, such as geographical proximity or lexical overlap, which could lead to better optimized arrangements of the crosslingual space.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462. ACL.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

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

Fabienne Braune, Viktor Hangya, Tobias Eder, and Alexander Fraser. 2018. Evaluating bilingual word embeddings on the long tail. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 188–193.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270.

Bernard Comrie. 1989. *Language Universals and Linguistic Typology: Syntax and Morphology*, University of Chicago Press.

Alexis Conneau, Guillaume Lample, Marc' Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, fast cross-lingual word-embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1113.

Paula Czarnowska, Sebastian Ruder, Édouard Grave, Ryan Cotterell, and Ann Copestake. 2019. Don't forget the long tail! A comprehensive analysis of morphological generalization in bilingual lexicon induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNP)*, pages 973–982.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*.

Yerai Doval, Jose Camacho-Collados, Luis Espinosa Anke, and Steven Schockaert. 2018. Improving cross-lingual word embeddings by meeting in the middle. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 294–304. ACL.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295.

Goran Glavas, Robert Litschko, Sebastian Ruder, and Ivan Vulic. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. *arXiv preprint arXiv:1902.00508*.

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

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Karl Moritz Hermann and Phil Blunsom. 2013. Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173*.

Geert Heyman, Bregt Verreet, Ivan Vulić, and Marie Francine Moens. 2019. Learning unsupervised multilingual word embeddings with incremental multilingual hubs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1890–1902.

Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2019. Learning multilingual word embeddings in latent metric space: a geometric approach. *Transactions of the Association for Computational Linguistics*, 7:107–120.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. Panlex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 3145–3150.

Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. 2018. Generalizing procrustes analysis for better bilingual dictionary induction. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 211–220.

Stanislas Lauly, Alex Boulanger, and Hugo Larochelle. 2014. Learning multilingual word representations using a bag-of-words autoencoder. *arXiv preprint arXiv:1401.1803*.

M. Paul Lewis and F. Gary. 2015. Simons, and Charles D. Fennig (eds.). 2013. *Ethnologue: Languages of the world,* pages 233–62.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174. ACL.

Robert Litschko, Goran Glavaš, Ivan Vulic, and Laura Dietz. 2019. Evaluating resource-lean cross-lingual embedding models in unsupervised retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1109–1112. ACM.

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A survey of cross-lingual word embedding models. *arXiv preprint arXiv:1706.04902*.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. 2013. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.

Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. 2015. Ridge regression, hubness, and zero-shot learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 135–151. Springer.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*.

Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 2015. Inverted indexing for cross-lingual NLP. In *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4398–4409.

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

Takashi Wada, Tomoharu Iwata, and Yuji Matsumoto. 2019. Unsupervised multilingual word embedding with limited resources using neural language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3113–3124.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.