NORMA: Neighborhood Sensitive Maps for Multilingual Word Embeddings

Ndapa Nakashole
Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093
nnakashole@eng.ucsd.edu

Abstract

Inducing multilingual word embeddings by learning a linear map between embedding spaces of different languages achieves remarkable accuracy on related languages. However, accuracy drops substantially when translating between distant languages. Given that languages exhibit differences in vocabulary, grammar, written form, or syntax, one would expect that embedding spaces of different languages have different structures especially for distant languages. With the goal of capturing such differences, we propose a method for learning neighborhood sensitive maps, NORMA. Our experiments show that NORMA outperforms current state-of-the-art methods for word translation between distant languages.

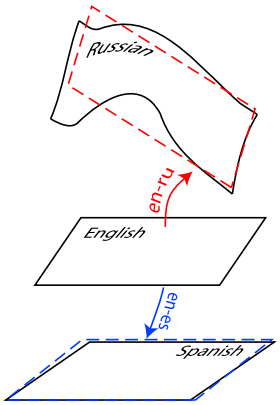

Figure 1: **Bottom:** By learning a linear map between embedding spaces of related languages, e.g., en-es, current methods achieve high accuracy on word translation. **Top:** For distant language pairs, e.g., en-ru, where differences are larger, word translation accuracy substantially degrades.

1 Introduction

The success of monolingual word embeddings has sparked interest in multilingual word embeddings. The goal is to learn word vectors where similar words have similar vector representations regardless of their language. Multilingual word embeddings are playing an increasingly prominent role in machine translation (Zou et al., 2013; Lample et al., 2018; Artetxe et al., 2018b). In addition, they are a promising avenue for cross-lingual model transfer (Guo et al., 2015; Täckström et al., 2012).

A prominent approach to learning multilingual word embeddings is to induce a mapping function between embedding spaces of different languages. However, there is a key assumption behind learning such a mapping function: that the embedding spaces of different languages exhibit similar structures (Mikolov et al., 2013a). Evidence that this assumption holds has mostly been through extrinsic evaluation metrics such as word translation accuracy. A notable exception is (Mikolov et al., 2013a), who showed empirical evidence on animals and numbers. Embeddings corresponding to a few numbers and animals in English and Spanish were projected down to two dimensions using PCA, and then manually rotated to accentuate similarity. Despite showing only these two concepts for two related languages, this work concluded that embedding spaces of different languages exhibit similar geometric arrangements. Additionally, work in this line of inquiry has continued to develop methods based on this assumption (Artetxe et al., 2018a; Conneau et al., 2018). Given that languages differ along dimensions such as vocabulary, grammar, written form, and syntax, one would expect that embedding spaces of different languages exhibit different structures. Indeed, recent work showed that assumptions of isomorphism and linearity do not hold (Søgaard et al., 2018; Nakashole and Flauger, 2018)

While these assumptions do not substantially affect accuracy when translating between related

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 512–522
Brussels, Belgium, October 31 - November 4, 2018.

languages, this is not the case for distant languages, see Figure 1. There is no established quantitative metric for measuring distances between languages. Language trees trace the evolution of languages but do not provide similarity scores. (Chiswick and Miller, 2005) learned similarity scores of 43 different languages to English by measuring how well Americans could learn a given language in a fixed period of time. Low scores on a standardized proficiency test were taken to indicate a large distance between the language and English. According to their scores, Japanese and Chinese are the most distant from English, Russian has a middle score, and French, Portuguese, Dutch, as expected, have some of the highest scores.

Additionally, linguists and psychologists have long studied the question of how language affects the way we think (Birner, 1999; Boroditsky, 2011). This influence would arise due to different languages organizing concepts differently.

We would like to model some aspects of the structural differences of languages when learning mapping functions between embedding spaces. To this end, we propose to learn neighborhood sensitive maps. We can, in principle, achieve neighborhood sensitive maps by training non-linear functions. However, training non-linear functions, in particular deep neural networks for this problem is difficult to optimize for this zero-shot (Lazaridou et al., 2015) learning problem, as we show in our experiments. Prior work alludes to similar observations(Mikolov et al., 2013a). For example, (Conneau et al., 2018) found that using non-linear mapping functions made training unstable[1].

In summary, our contributions are as follows:

- We propose a method for learning neighborhood sensitive maps, NORMA, which learns a single mapping function but in a departure from prior work, it *discovers neighborhoods*. NORMA avoids learning multiple mapping functions, thus enabling parameter sharing among neighborhoods. This is a more efficient use of training data than if we were to train multiple mapping functions for different neighborhoods as is done in (Zou et al., 2013).

- The neighborhoods are learned jointly while learning to translate, and we show that they are *interpretable*.

- Our experiments show that for word translation between distant languages, NORMA substantially outperforms methods that achieve the best performance when translating between related languages.

- Additionally, in the related language setting, we show that on rare words NORMA substantially outperforms state-of-the-art methods.

[1] https://openreview.net/forum?id=H196sainb

2 Related Work

The common approach to learning cross embedding space mapping functions is: first monolingual word embeddings for each language are trained independently; and second, a mapping function is learned, using supervised or unsupervised methods. The resulting mapping function enables translating words from the source to the target language.

Map Induction Methods. The earliest and simplest approach is to use a regularized least squares loss to induce a linear map $\mathbf{M}$ as follows: $\hat{\mathbf{M}} = \arg\min_{\mathbf{M}} ||\mathbf{MX} - \mathbf{Y}||_F + \lambda||\mathbf{M}||$, here $\mathbf{X}$ and $\mathbf{Y}$ are matrices that contain word embedding vectors for the source and target language (Mikolov et al., 2013a; Dinu et al., 2014; Vulic and Korhonen, 2016). Improved results were obtained by imposing an orthogonality constraint on $\mathbf{M}$ (Xing et al., 2015; Smith et al., 2017). Another loss function used in prior work is the max-margin loss, which has been shown to significantly outperform the least squares loss (Lazaridou et al., 2015; Nakashole and Flauger, 2017).

Another approach is to use canonical correlation analysis (CCA) to map two languages to a shared embedding space (Haghighi et al., 2008; Faruqui and Dyer, 2014; Lu et al., 2015; Ammar et al., 2016).

Most of the prior methods can be characterized as a series of linear transformations. In particular, (Artetxe et al., 2018a) propose a framework to differentiate prior methods in terms of which transformations they perform: embedding normalization, whitening, re-weighting, de-whitening, and dimensionality reduction.

Work on phrase translation proposed to induce many local maps that are individually trained (Zhao et al., 2015) on local neighborhoods. In

contrast, our approach trains a single function while taking into account neighborhood sensitivity. Our underlying motivation of neighborhood sensitivity is similar in spirit to the use of locally linear embeddings for nonlinear dimensionality reduction (Roweis and Saul, 2000).

Forms of Supervision. The methods we have described so far fall under supervised learning. In the supervised setting, a seed dictionary (5k word pairs is a typical size) is used to induce the mapping function. In (Artetxe et al., 2017) a semi-supervised approach is explored, whereby the method alternates between learning the map and generating an increasingly large dictionary. Completely unsupervised methods have recently been proposed using adversarial training (Barone, 2016; Zhang et al., 2017; Conneau et al., 2018). However, the underlying methods for learning the mapping function are similar to prior work such as (Xing et al., 2015). The limitations and strengths of unsupervised methods are detailed in (Søgaard et al., 2018)

Although in our our experiments we work in the supervised setting, NORMA can work with any form of supervision.

Translation Retrieval Methods. The most commonly used way to obtain a translation t of a source language word s is nearest neighbor retrieval, given by: $t = \arg\max_t \cos(\mathbf{M}x_s, y_t)$. Alternative retrieval methods have been proposed, such as the inverted nearest neighbor retrieval(Dinu et al., 2014), inverted softmax (Smith et al., 2017) and Cross-Domain Similarity Local Scaling (CSLS) (Conneau et al., 2018). Since we are interested in evaluating the quality of mapping functions, our experiments use standard nearest neighbor retrieval for all methods.

3 Local Maps in Embedding Space

Is it useful for maps to be neighborhood sensitive? To study this question we carried out experiments comparing performance of neighborhood-specific maps to global maps. A thorough analysis of this kind was carried out in our prior work (Nakashole and Flauger, 2018)

We created neighborhoods by first selecting the embeddings of a few words associated with specific topics such as diseases, or cities. We then added all nearby words, which are words whose cosine similarity to any of the selected words is >=0.5 [2]. We used three language pairs for local vs global map translation experiments: English to German, English to Portuguese, and English to Swedish. The neighborhoods and their train/test splits are:
$en - de$: medication(3,415/500), cities(2,083/500), and animals(990/500);
$en - pt$: diseases(1,670/300), chemicals(1,279/300), and names(1,986/300);
$en - sv$: flowers(1,537/200), insects(1,271/200), and names(1,416/200). The training and test data was obtained from subsets of Facebook AI MUSE lexicons[3]

For each of the neighborhoods, we evaluated translation accuracy both when using a locally trained map and when using a globally trained map. The difference is that the locally trained map is only trained using training data from the neighborhood, whereas the global map is trained using training data from the neighborhood but also from all other neighborhoods and more (˜10000 word pairs). That is, the training data for global maps is a superset of the local training data.

We trained all maps using linear transformations. As we will show in our experiments, optimizing neural network mapping functions for this problem fails. This is a similar observation to prior work (Mikolov et al., 2013a; Conneau et al., 2018)[1]. More details on models and experimental settings are described in Sections 4 and 5.

Figure 2 shows that for various neighborhoods, translation accuracy is higher when we train neighborhood-specific maps than one single global map. These results are similar to (Zou et al., 2013) who then trained many local maps. While we could also proceed to train many local maps, this requires identifying optimal neighborhoods. It also requires gathering sufficient training data for each of the neighborhoods independently. In our proposed method, NORMA, we avoid learning multiple maps, creating a single map, while modeling neighborhood information and promoting parameter sharing.

Overall, the results in Figure 2 are an indicator neighborhood sensitivity in maps is useful. This would particularly be useful for distant languages

[2]We found a 0.5 cutoff to be a good compromise between neighborhood purity, and size. However, our final method (Section 4) on which all our comparison experiments were based, automatically discovers neighborhoods based on ideas from sparse coding.

[3]https://github.com/facebookresearch/MUSE

where a single global map that is linear might not suffice since the underlying embedding structure for distant languages might differ more than those of related languages as depicted in Figure 1.

4 Model

In this section we introduce our model for learning neighborhood sensitive maps, NORMA. Our approach jointly discovers neighborhoods while learning to translate.

4.1 Reconstructive Neighborhood Discovery

Inspired by work on sparse coding (Lee et al., 2007), we discover neighborhoods by learning a reconstructive dictionary. We would like to learn a dictionary of neighborhoods on the source language side. To learn this dictionary, we set up a reconstruction objective, where for any given word embedding $x_i \in \mathbb{R}^d$, where d is the dimensionality of the word embeddings, we want to reconstruct x_i using a linear combination of K neighborhoods. Let $\mathbf{D} \in \mathbb{R}^{K \times d}$ be the neighborhood matrix, each row of $\mathbf{D}$ represents a d-dimensional vector which can be interpreted as representing the center of the neighborhood. Let $\mathbf{X} \in \mathbb{R}^{N \times d}$ be a set of N embedding vectors corresponding to words in the source language vocabulary[4]. We can learn a reconstructive dictionary of K neighborhoods with the following objective:

$$\mathbf{D}, \mathbf{V} = \arg\min_{\mathbf{D},\mathbf{V}} ||\mathbf{X} - \mathbf{V}\mathbf{D}||_2^2 \quad (1)$$

$\mathbf{D} \in \mathbb{R}^{K \times d}$ is the learned dictionary of neighborhoods, $K > d$ and thus the dictionary is overcomplete; $\mathbf{V} \in \mathbb{R}^{N \times K}$ are the learned neighborhood membership weights for $\mathbf{X}$. While we use the squared loss, other loss functions can be used (Lee et al., 2007). To encourage neighborhoods to be different from each other, one can impose an orthogonality constraint : $||\mathbf{D}\mathbf{D}^T - \mathbf{I}||$ where $\mathbf{I}$ is the identity matrix. The reconstruction error with an orthogonality penalty is:

$$R(\theta) = ||\mathbf{X} - \mathbf{V}\mathbf{D}||_2^2 + \lambda||\mathbf{D}\mathbf{D}^T - \mathbf{I}|| \quad (2)$$

Where λ is a hyperparameter which controls the contribution of the orthogonality constraint to the reconstruction error.

[4]Since the vocabulary size can be very large, in our experiments, we work in batches of N=50

4.2 Joint Neighborhood Discovery and Translation

Our approach ties neighborhood discovery to the word translation task. First, we obtain neighborhood 'factorized' representations by multiplying the input vector $\mathbf{X}$ by the dictionary of neighborhoods:

$$\mathbf{X}_{\boldsymbol{\mathcal{N}}} = \mathbf{X}\mathbf{D}^T,$$

where $\mathbf{X}_{\boldsymbol{\mathcal{N}}} \in \mathbb{R}^{N \times K}$. Here again N refers to words in the source language vocabulary, English in the case of $en - de$ translation. And K is the number of neighborhoods.

Second, we obtain an intermediate representation of the input, which contains both the original input $\mathbf{X}$ and the neighborhood 'factorized' representations of the input $\mathbf{X}_{\boldsymbol{\mathcal{N}}}$, through vector concatenation as follows:

$$\mathbf{X}_{\boldsymbol{\mathcal{I}}} = [\mathbf{X}_{\boldsymbol{\mathcal{N}}}; \mathbf{X}],$$

where $\mathbf{X}_{\boldsymbol{\mathcal{I}}} \in \mathbb{R}^{N \times (K+d)}$.

To get the final representation of the input, we project $\mathbf{X}_{\boldsymbol{\mathcal{I}}}$ into a low-dimensional vector of the same size as the original input:

$$\mathbf{X}_{\boldsymbol{\mathcal{F}}} = \mathbf{X}_{\boldsymbol{\mathcal{I}}}\mathbf{W}_f,$$

where $\mathbf{W}_f \in \mathbb{R}^{(K+d) \times d}$ is a set of learned parameters. And $\mathbf{X}_{\boldsymbol{\mathcal{F}}} \in \mathbb{R}^{N \times d}$ is the resulting final representation.

We use these neighborhood sensitive representation $\mathbf{X}_{\boldsymbol{\mathcal{F}}}$ as the input for learning the mapping function $\mathbf{W}$, instead of the original $\mathbf{X}$. We explore different ways for learning the mapping $\mathbf{W}$: first a linear mapping, and second, a single layer neural network with a leaky rectified linear unit (leaky ReLU[5]) non-linearity and a highway layer (Srivastava et al., 2015). As we will show in our experiments, training neural networks with more layers fails on this zero-shot learning problem.

For the linear map, the translation $\hat{y}_i$ is given by:

$$\hat{y}_i^{linear} = \mathbf{W}x_{\mathcal{F}_i} \quad (3)$$

where $x_{\mathcal{F}_i} \in \mathbf{X}_{\boldsymbol{\mathcal{F}}}$ is the neighborhood sensitive representation of x_i.

For the neural network map, using a single layer neural network, and a highway layer, the transla-

[5]It outperformed other non-linearities such as $tanh$ in our initial experiments.

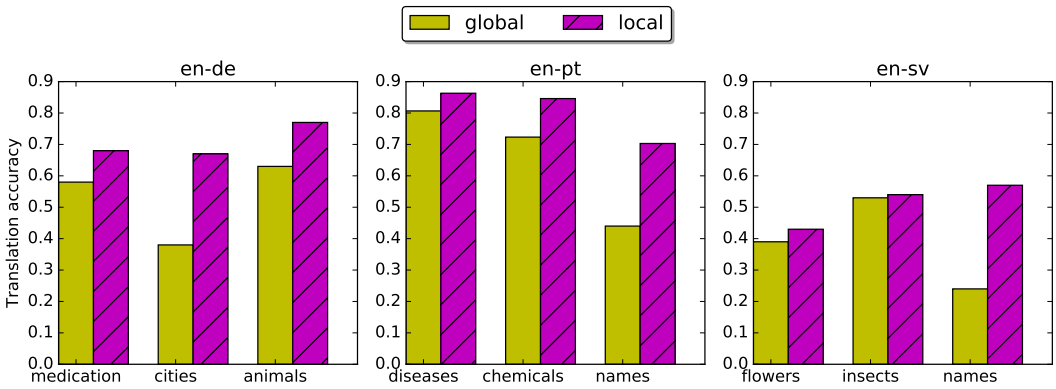

Figure 2: Accuracy of globally vs locally trained mapping functions for various neighborhoods on $en-de$, $en-pt$, and $en-sv$ translation.

tion $\hat{y}_i$ is given by:

$$h_i = \sigma_1(x_{\mathcal{F}_i}\mathbf{W})$$
$$t_i = \sigma_2(x_{\mathcal{F}_i}\mathbf{W^t})$$
$$\hat{y}_i^{nn} = t_i \times h_i + (1.0 - t_i) \times x_{\mathcal{F}_i} \quad (4)$$

where σ_1 is a non-linearity. We use a leaky-ReLU non-linearity. σ_2 is the sigmoid function. $\mathbf{W^t}$ is another set of parameters in addition to $\mathbf{W}$.

4.3 Objective Function

We use the max-margin loss function to learn the parameters of the model:

$$L(\theta) = \sum_{i=1}^{m}\sum_{j\neq i}^{k} \max\Big(0, \gamma + d(y_i, \hat{y}_i^g) - d(y_j, \hat{y}_i^g)\Big), \quad (5)$$

Where y_i is the true label; $\hat{y}_i^g$ is the prediction, which is either $\hat{y}_i^{linear}$ or $\hat{y}_i^{nn}$. The goal of the max-margin loss function is to rank correct training data pairs (x_i, y_i) higher than incorrect pairs (x_i, y_j) with a margin of at least γ. The margin γ is a hyper-parameter and the incorrect labels, y_j are selected randomly such that $j \neq i$. k is the number of incorrect examples per training instance, and $d(x, y) = (x-y)^2$ is the distance measure.

The joint neighborhood discovery and word translation objective is given by:

$$J(\theta) = L(\theta) + R(\theta) \quad (6)$$

The neighborhood discovery part of the objective, $R(\theta)$, does not depend on availability of supervised data and only requires monolingual data on the source language side. Thus, we can discover neighborhoods in an unsupervised manner on a large set of monolingual word embeddings, then initialize using this pre-trained D which is then jointly optimized with the translation part of the objective $L(\theta)$. Importantly, this also means that our method can work with unsupervised methods for learning mapping functions such as those using adversarial training (Barone, 2016; Conneau et al., 2018).

5 Experimental Evaluation

In this section, we study the following questions: How does NORMA compare to state-of-the-art methods for learning mapping functions between embedding spaces of different languages? We study this question in three settings: when translating between distant languages, when translating between related languages, and lastly, when translating between related languages but on rare words. Additionally, we ask the following question: are the neighborhoods learned by NORMA meaningful?

To study these questions, we carried out experiments on word translation from English to two distant languages, a Slavic language (Russian), and a Sino-Tibetan language (Chinese). In addition, we carried out experiments on word translation between related languages (English, French, German and Portuguese).

Data and Experimental Setup. The Facebook AI MUSE[3] project (Conneau et al., 2018) provides train/test data for bilingual dictionaries of various language pairs, we use this data in our experiments. The MUSE dictionaries consist of

Method	Slavic & Sino-Tibetan				
	en-ru	en-zh	en-de	en-es	en-fr
NORMA-Linear	**50.33**	**43.27**	68.50	77.47	76.10
NORMA-Highway-NN	49.27	33.10	67.33	77.65	75.50
1 layer-NN	49.13	30.66	66.80	77.60	75.53
2 layer-NN	0	0	0	0	0
1 layer-Highway-NN	49.50	30.91	67.00	77.50	75.60
2 layer-Highway-NN	0	0	0	0	0
Artetxe et al . 2018	47.93	20.4	70.13	**79.6**	**79.30**
Conneau et al. 2018	37.30	30.90	**71.30**	79.10	78.10
Smith et al. 2017	46.33	39.60	69.20	78.80	78.13
Xing et al. 2015	44.50	41.0	67.07	77.33	75.47
Lazaridou et al. 2015	48.27	29.60	68.20	77.60	75.86
Faruqui and Dyer (2014)	35.47	32.20	55.67	72.33	69.27
Mikolov et al. 2013	42.47	19.80	60.07	74.20	71.60

Table 1: Precision at 1 comparison of NORMA to previously proposed mapping functions. We used FAIR/MUSE word translation lexicons train/test splits.

	en-ru	en-zh	en-de	en-es	en-fr
NOUN	42% /**55.1**	42% /42.1	39% /**74.6**	40% /**82.3**	42% /**80.0**
VERB	41% /47.3	39% /**47.6**	38% /64.4	40% /71.6	41% /70.0
ADJECTIVE	10% /34.4	11% /38.7	10% /56.1	9% /76.9	10% /71.3

Table 2: Part-of-Speech (POS) distributions of the MUSE test sets. Listed are the top 3 parts of speech, which account for ˜90% of the test data for all language pairs. X% /Y means the POS tag makes up X% of the test set, with accuracy Y.

5,000/1,500 word pairs for train/test data. Unless specified, we use the train/test split provided by MUSE. Development sets: the MUSE dictionaries that we used are very large. They contain over 100,000 entries for most language pairs, we tuned our models on data that was not part of the train and test sets.

We obtained pre-trained word embeddings from FastText (Bojanowski et al., 2017). In Equation 2, we did not find it helpful to encourage neighborhoods to be different, thus we set $\lambda = 0$. We set the margin γ in Equation 5 to be $\gamma = 0.4$. For the dictionary of neighborhoods D in Equation 1, we set the number of neighborhoods $K = 2,000$[6]. We use $N = 50$ batch size. We estimate model parameters using stochastic gradient descent.

Methods Under Comparison. We compare variations of NORMA to several previously proposed methods for generating mapping functions. The methods compared are: (Artetxe et al., 2018a; Conneau et al., 2018; Smith et al., 2017; Xing et al., 2015; Lazaridou et al., 2015; Faruqui and Dyer, 2014; Mikolov et al., 2013a). More detailed descriptions of these prior methods can be found in the related work section.

Our primary goal is to evaluate the quality of maps produced. While a number of prior work proposed various approaches for retrieval, which have been shown to improve accuracy by a few points, we compare all methods using the same retrieval method, nearest neighbor. Thus, for (Conneau et al., 2018), we report the results for the variant of their method called: *adv - Refine - NN*.

[6]We carried out experiments using different neighborhood sizes, and consistently found $K \approx 2000$ to outperform other choices.

5.1 English to Slavic and Sino Tibetan

State-of-the-art methods have mostly focused word translation evaluation on English to Latin languages or other nearby languages. (Artetxe et al., 2018a) performed experiments on en-es, en-de, en-it and en-fi, where concepts might still be organized in a relatively similar way. In (Conneau et al., 2018), the adversarial training method proposed was evaluated on Chinese, Russian, and Esperanto, but thorough comparison experiments to prior work on word translation were only performed on English to Italian.

We carried out en-ru and en-zh comparison experiments, and present the results in the second and third columns of Table 1. The two state-of-of-the art methods (Artetxe et al., 2018a) and (Conneau et al., 2018) are significantly outperformed by NORMA-Linear. On English to Russian, NORMA-Linear achieves 50.33 precision 1, outperforming both (Artetxe et al., 2018a) (Conneau et al., 2018), as well as other methods. On English to Chinese, NORMA-Linear achieves 43.37 precision 1, again ahead of other methods. The best performing variant of our method is NORMA-Linear. The neural networks with more than a single layer prove difficult to optimize for this problem, and produce accuracy of 0. This could be because the problem of cross-embedding space mapping is a zero-shot learning problem, which is much more difficult to train than a supervised problem, the setting in which deep learning methods have thrived so far.

5.2 English to Related Languages

We show experiments on English to related languages in the last three columns of Table 1. On these languages, indeed the most recently proposed methods (Artetxe et al., 2018a; Conneau et al., 2018) produce the best performing maps. However, NORMA-Linear is only 2-3 points behind these methods. This in contrast to English to Chinese where both (Artetxe et al., 2018a) and (Conneau et al., 2018) are behind NORMA - Linear, by more than 10 points.

A promising line of future work is to get NORMA-Linear to bridge the 2-3 point gap on related languages by exploring a best of both worlds approach, combining neighborhood sensitivity with the methods that achieve superior performance on nearby languages.

	en-pt	
	RARE	MUSE
NORMA-Linear	**57.67**	72.60
NORMA-Highway-NN	49.33	71.73
1 layer-NN	48.67	72.13
1 layer-Highway-NN	49.33	72.10
Artetxe et al . 2018	47.00	**77.73**
Lazaridou et al 2015	48.00	72.27

Table 3: Performance for en-pt on rare words (RARE), and the en-pt MUSE dataset, which as shown in Figure 3 contains a lot of frequent words.

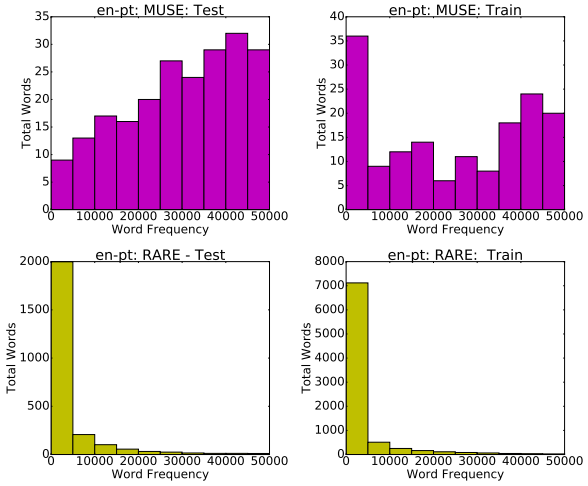

Figure 3: Top: Frequency distribution of MUSE dictionary test and train sets for en-pt. Bottom: Frequency distribution of the RARE words dataset.

5.3 Accuracy by Part-of-Speech

We assigned each word its majority part-of-speech by tagging the ClueWeb[7] corpus, which contains over 500 million webpages. We then evaluated translation precision of NORMA-Linear stratified by part-of-speech. The results are shown in Table 5 We found that, nouns and verbs make up about 80 percent of the MUSE test dictionaries, followed by adjectives (˜10%). We found that while nouns and verbs make up a large chunk of the test data, nouns are translated with much higher accuracy than verbs, except for English to Chinese. This finding will serve as a guide for future improvements to our method.

5.4 English to Languages: Rare Words

We analyzed the frequency distribution of the MUSE dictionaries. To get word frequency infor-

[7] https://www.lemurproject.org/clueweb09.php/

Neighborhood			
51	134	162	7
drugs	criminally	chuanyao	khoisan
zonisamide	judicature	chuanyan	bantu
cocaine	prosecutory	zhiang	sepedi
ritalin	derogation	thanong	otjiherero
hospitalized	restitutionary	qiangbing	ndebeles
pheniprazine	derogative	pengpeng	hereros
overdose	jailable	nguyan	otjinene
disorientation	extradition	yuning	shona
focusyn	sodomy	liheng	hutu
alfaxalone	crimes	thanong	witotoan

Table 4: Sample neighborhoods discovered by NORMA during en-de translation: 51 appears to represent drugs, 132: justice and crime; 162: Asian names, 7 : African names.

mation, we processed documents in the ClueWeb[7] corpus and recorded word occurrence frequency. We discovered that the MUSE dictionaries contain a lot of frequent words. The top half of Figure 3 shows frequency counts of the en-pt MUSE test dictionary. For readability we only show bins up to occurrence frequency of 50,000. We see that only about 50/1500 in the MUSE en-pt test data are infrequent, the rest are frequent words, occurring more than 10,000 times in the ClueWeb corpus.

We therefore created another test set for en-pt from the rest of the MUSE data which is not part of the train or test data, with the goal of creating a train/test of rare words. The bottom half of Figure 3 is a plot of frequency counts of train and test data for these rare words.

We then compared variations of NORMA to the best performing method on English to related languages, which is (Artetxe et al., 2018a). The comparison was done both on the regular MUSE test dataset for en-pt and the rare word dataset for en-pt. Since our method uses a max-margin loss much like (Lazaridou et al., 2015), we also compare to (Lazaridou et al., 2015).

Table 3 shows that NORMA-Linear outperforms (Artetxe et al., 2018a) by over 10 points on the RARE words dataset. On the regular MUSE dictionary, (Artetxe et al., 2018a) is ahead by about 5 points. On RARE, (Lazaridou et al., 2015) is behind NORMA-Linear by 9 points, whereas on the MUSE dictionary performance of (Lazaridou et al., 2015) and NORMA-Linear is about the same.

5.5 Neighborhood Interpretability

NORMA jointly discovers neighborhoods while learning to translate words. We now ask if the discovered neighborhoods semantically make sense. We can answer this question since each neighborhood vector can be seen as a "center" vector representing the words in the neighborhood. Thus we can consider words whose cosine similarity to the neighborhood vector is greater than some threshold, to be members of that neighborhood. As we mentioned, we found that setting the total number of neighborhoods to be discovered to $K = 2,000$ provided the best results. Of these 2,000 we show some of them in Table 4 obtained when training $en - de$. For each neighborhood, we show 10 words that appear among the top 100 words of that neighborhood. It can be seen that the neighborhoods represent some kind of "topics". For example, neighborhood number 51 appears to represent drugs, and drug-related concepts; number 132 contains justice and crime-related concepts; number 162 contains mostly Asian concepts and names, number 7 contains mostly African and names. We can see that the granularity of neighborhoods and their specificity varies.

6 Conclusions

We propose neighborhood sensitive maps for learning multilingual word embeddings, NORMA. Our method is motivated by the fact that languages differ along dimensions such as vocabulary, grammar, written form, and syntax, and therefore one would expect that embedding spaces of different languages exhibit different structures especially for distant languages.

Our method jointly discovers neighborhoods while learning to translate words. Experimental evaluation showed that NORMA substantially outperforms state-of-the-art (SOTA) methods on distant languages, while only being a few points behind on related languages. A promising line of future work is to explore a best of both worlds approach, combining neighborhood sensitivity with the methods that achieve superior performance on nearby languages.

Acknowledgments

We thank Raphael Flauger for useful discussions, and the anonymous reviewers for their constructive comments.

References

Hanan Aldarmaki, Mahesh Mohan, and Mona Diab. 2018. Unsupervised word mapping using structural similarities in monolingual embeddings. *Transactions of the Association of Computational Linguistics*, 6:185–196.

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

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)*, volume 1, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *AAAI*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *ICLR*.

Antonio Valerio Miceli Barone. 2016. Towards crosslingual distributed representations without parallel text trained with adversarial autoencoders. In *1st Workshop on Representation Learning for NLP*.

Betty Birner. 1999. Does the language i speak influence the way i think?.

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL*.

Lera Boroditsky. 2011. How language shapes thought. *Scientific American*, 304(2):62–65.

A. P. Sarath 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.

Barry R Chiswick and Paul W Miller. 2005. Linguistic distance: A quantitative measure of the distance between english and other languages. *Journal of Multilingual and Multicultural Development*, 26(1):1–11.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data.

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

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

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

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *NAACL*, pages 1386–1390.

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.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *ACL*, pages 771–779.

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

Tomáš Kočiskỳ, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. *arXiv preprint arXiv:1405.0947*.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *ACL Workshop on Unsupervised Lexical Acquisition*.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *ICLR*.

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.

Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng. 2007. Efficient sparse coding algorithms. In *NIPS*, pages 801–808.

Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong. 2014. Learning small-size dnn with output-distribution-based criteria. In *INTERSPEECH*, pages 1910–1914.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256.

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, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195.

Ndapa Nakashole and Raphael Flauger. 2018. Characterizing departures from linearity in word translation. In *ACL*.

Ndapandula Nakashole and Raphael Flauger. 2017. Knowledge distillation for bilingual dictionary induction. In *EMNLP*, pages 2487–2496.

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

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *ACL*.

Sam T Roweis and Lawrence K Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326.

Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. 2015. Learning cross-lingual word embeddings via matrix co-factorization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 567–572.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *ICLR*.

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

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks. *arXiv preprint arXiv:1505.00387*.

Oscar Täckström, Ryan T. McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *NAACL*, pages 477–487.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res. (JAIR)*, 37:141–188.

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

Ivan Vulic and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. ACL.

Ivan Vulic and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *ACL*, pages 719–725.

Derry Tanti Wijaya, Brendan Callahan, John Hewitt, Jie Gao, Xiao Ling, Marianna Apidianaki, and Chris Callison-Burch. 2017. Learning translations via matrix completion. In *EMNLP*, pages 1452–1463.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *HLT-NAACL*, pages 1006–1011.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1959–1970.

Kai Zhao, Hany Hassan, and Michael Auli. 2015. Learning translation models from monolingual continuous representations. In *NAACL*, pages 1527–1536.

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.