Multi-lingual Common Semantic Space Construction via Cluster-consistent Word Embedding

Lifu Huang[1], **Kyunghyun Cho**[2], **Boliang Zhang**[1], **Heng Ji**[1], **Kevin Knight**[3]
[1] Rensselaer Polytechnic Institute
{huangl7,zhangb8,jih}@rpi.edu
[2] New York University, CIFAR Global Scholar
kyunghyun.cho@nyu.edu
[3] Didi Labs and University of Southern California
knight@isi.edu

Abstract

We construct a multilingual common semantic space based on distributional semantics, where words from multiple languages are projected into a shared space via which all available resources and knowledge can be shared across multiple languages. Beyond word alignment, we introduce multiple cluster-level alignments and enforce the word clusters to be consistently distributed across multiple languages. We exploit three signals for clustering: (1) neighbor words in the monolingual word embedding space; (2) character-level information; and (3) linguistic properties (e.g., apposition, locative suffix) derived from linguistic structure knowledge bases available for thousands of languages. We introduce a new cluster-consistent correlational neural network to construct the common semantic space by aligning words as well as clusters. Intrinsic evaluation on monolingual and multilingual QVEC tasks shows our approach achieves significantly higher correlation with linguistic features which are extracted from manually crafted lexical resources than state-of-the-art multi-lingual embedding learning methods do. Using low-resource language name tagging as a case study for extrinsic evaluation, our approach achieves up to 14.6% absolute F-score gain over the state of the art on cross-lingual direct transfer. Our approach is also shown to be robust even when the size of bilingual dictionary is small.[1]

1 Introduction

More than 3,000 languages have electronic record, e.g., at least a portion of the Christian Bible had been translated into 2,508 different languages. However, the training data for mainstream natural language processing (NLP) tasks such as Information Extraction (IE) and Machine Translation (MT) is only available for dozens of dominant languages. In this paper we aim to construct a multilingual common semantic space where words in multiple languages are mapped into a distributed, language-agnostic semantic continuous space, so that resources and knowledge can be shared across languages.

Previous multilingual embedding methods align the semantic distributions of words from multiple languages within the common semantic space. Though several recent attempts (Artetxe et al., 2017, 2018; Conneau et al., 2017) have shown that it is possible to extract multilingual word embedding from a pair of potentially unaligned corpora in multiple languages, we claim that it is necessary to impose more constraints to preserve linguistic properties and facilitate downstream NLP tasks, such as cross-lingual IE, and MT. We find that words also can be clustered through explicit (e.g., sharing affixes of certain linguistic functions) or implicit clues (e.g., sharing neighbors from monolingual word embedding) and such clusters should also be consistent across multiple languages. To do so, we design a new algorithm, called cluster-consistent multilingual word embedding, that extracts multilingual word embedding vectors which preserve the natural clustering structures of words across multiple languages.

We propose to create clusters through three kinds of signals as follows, without any extra human annotation effort. Then we aggregate the embedding vectors of words in each cluster and ensure that the clusters (or the words therein) are consistent across multiple languages.

Neighbor based clustering and alignment. We build our common space based on correlational neural network (CorrNet) which is an extension of autoencoder framework by enabling cross-lingual reconstruction. In contrast to previous work (Chandar et al., 2016; Rajendran et al.,

[1]The resources and programs are available for research purpose: https://github.com/wilburOne/CommonSpace/

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 250–260
Brussels, Belgium, October 31 - November 4, 2018.

2015), we extend CorrNet to *neighbor-consistent correlation network* by using each word's neighbors (the nearest words within monolingual semantic space) to ensure that the cross-lingual mapping from and to the common semantic space is locally smooth. For instance, the neighboring words of *China* in English (*Japan*, *India* and *Taiwan*) should be close to the neighboring words of *Cina* in Italian (*Beijing*, *Korea*, *Japan*) in the common semantic space. In other words, we encourage the consistency of neighborhoods across multiple languages.

Character based clustering and alignment. Many related languages share very similar character set, and many words that refer to the same concept share similar compositional characters or patterns, e.g., *China* (English), *Kina* (Danish), and *Cina* (Italian).

Linguistic property based clustering and alignment. Many languages also share linguistic properties, e.g., apposition, conjunction, and plural suffix (English (*-s* / *-es*), Turkish (*-lar* / *-ler*), Somali (*-o*)). Linguists have created a wide variety of linguistic property knowledge bases, which are readily available for thousands of languages. For example, the CLDR (Unicode Common Locale Data Repository)[2] includes closed word classes and affixes indicating various linguistic properties. We propose to take advantage of these language-universal resources to create clusters, where the words within one cluster share the same linguistic property, and build alignment between clusters for common semantic space construction.

We evaluate our approach on monolingual and multilingual QVEC (Tsvetkov et al., 2015) tasks, which measure the quality of word embeddings based on the alignment of the embeddings to linguistic feature vectors extracted from manually crafted linguistic resources, as well as an extrinsic evaluation on name tagging for low-resource languages. Experiments demonstrate that our framework is effective at capturing linguistic properties and significantly outperforms state-of-the-art multi-lingual embedding learning methods.

2 Related Work

Multilingual word embeddings have advanced many multilingual NLP tasks, such as machine translation (Zou et al., 2013; Mikolov et al., 2013b; Madhyastha and España-Bonet, 2017), dependency parsing (Guo et al., 2015; Ammar et al., 2016a), and name tagging (Zhang et al., 2017a; Tsai and Roth, 2016; Zhang et al., 2018; Cheung et al.; Zhang et al., 2017b; Feng et al., 2017). Using bilingual aligned words, previous methods project multiple monolingual embeddings into a shared semantic space using linear mappings (Mikolov et al., 2013b; Rothe et al., 2016; Zhang et al., 2016; Baroni et al., 2015; Xing et al., 2015; Smith et al., 2017) or canonical correlation analysis (CCA) (Ammar et al., 2016b; Faruqui and Dyer, 2014; Lu et al., 2015). Compared with CCA, which only optimizes the correlation for each individual pair of languages, linear mapping based methods can jointly optimize all the languages in the common semantic space. We focus on learning linear mappings to construct the common semantic space and adopt correlational neural networks (CorrNet) (Chandar et al., 2016; Rajendran et al., 2015) as the basic model. In contrast to previous work which only exploited monolingual word semantics, we introduce multiple cluster-level alignments and design a new cluster consistent CorrNet to align both words and clusters.

Another branch of approaches for multilingual word embeddings are based on parallel or comparable data, such as parallel sentences (AP Chandar et al., 2014; Gouws et al., 2015; Luong et al., 2015; Hermann and Blunsom, 2014; Schwenk et al., 2017), phrase translations (Duong et al., 2016) and comparable documents (Vulic and Moens, 2015). Moreover, to reduce the need of bilingual alignment, several approaches have been designed to learn cross-lingual embeddings based on a small seed dictionary (Vulic and Korhonen, 2016; Zhang et al., 2016; Artetxe et al., 2017), or even with no supervision (Cao et al., 2016; Zhang et al., 2017d,c; Conneau et al., 2017; Artetxe et al., 2018). However, such methods are still limited to bilingual word embedding learning and remaining to be explored for common semantic space construction.

3 Approach

3.1 Overview

Figure 1 shows the overview of our neural architecture. We project all monolingual word embeddings into a common semantic space based on word-level as well as cluster-level alignments and learn the transformation functions. First, on

[2] cldr.unicode.org

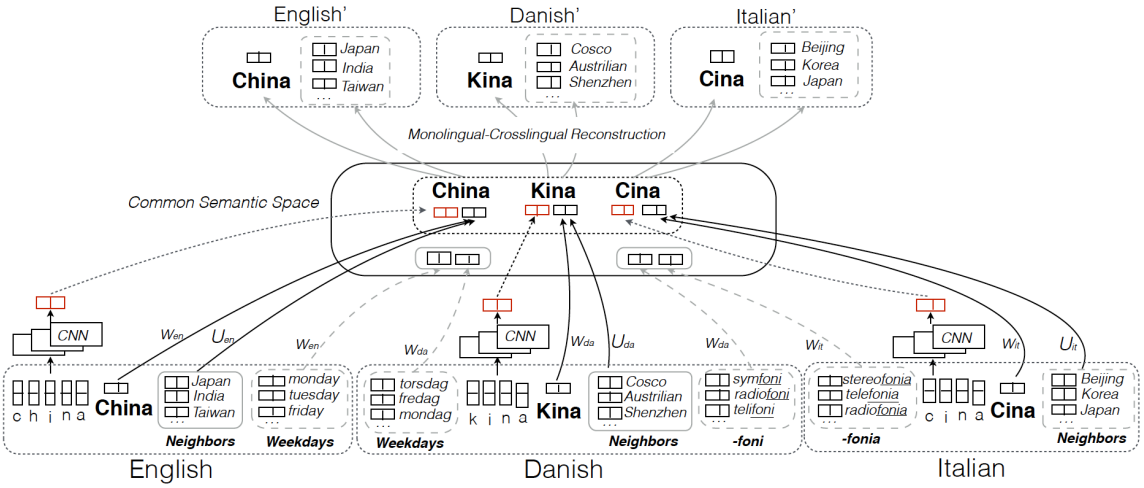

Figure 1: Architecture Overview. In each monolingual semantic space, the words within solid rectangle denote a neighbor based cluster and the words within dotted rectangle denote a linguistic property based cluster.

word-level, we build a neighborhood-consistent CorrNet to augment word representations with neighbor based clusters and align them in the common semantic space. In addition, we apply a language-independent convolutional neural networks to compose character-level word representation and concatenate it with word representation in the common semantic space. Finally, we construct clusters based on linguistic properties, including closed word classes and affixes, and align them in the common semantic space. We jointly optimize for all the alignments in the common semantic space for each pair of languages.

3.2 Basic Model

We briefly describe the basic model for learning the common semantic space: correlational neural networks (CorrNets) (Chandar et al., 2016; Rajendran et al., 2015). It combines the advantages of canonical correlation analysis (CCA) and autoencoder (AE).

Given the bilingual aligned word pairs between two languages l_1 and l_2, we first use their monolingual word embeddings to initialize each word with a vector and obtain $M_{l_1} \in \mathbb{R}^{|V_{l_1}| \times d_{l_1}}$ and $M_{l_2} \in \mathbb{R}^{|V_{l_2}| \times d_{l_2}}$, where V_{l_1} and V_{l_2} are the bilingual dictionary of l_1 and l_2. $V_{l_1}^i$ is the translation of $V_{l_2}^i$, and d_{l_1} and d_{l_2} are the vector dimensionalities. Then for each language we learn a linear projection function to project M_{l_1} and M_{l_2} into the common semantic space:

$$H_{l_1} = \sigma(M_{l_1} \cdot W_{l_1} + b_{l_1}) \,,$$

$$H_{l_2} = \sigma(M_{l_2} \cdot W_{l_2} + b_{l_2}) \,,$$

where $H_{l_1} \in \mathbb{R}^{|V_{l_1}| \times h}$ and $H_{l_2} \in \mathbb{R}^{|V_{l_2}| \times h}$ are the vector representations for V_{l_1} and V_{l_2} in the common semantic space respectively. h is the vector dimensionality in the shared semantic space. $W_{l_1} \in \mathbb{R}^{d_{l_1} \times h}$ and $W_{l_2} \in \mathbb{R}^{d_{l_2} \times h}$ are the transformation matrices, and b_{l_1} and b_{l_2} are the bias vectors. σ denotes Sigmoid function.

After we project the monolingual embeddings into the common semantic space, we further reconstruct M_{l_1} and M_{l_2} from H_{l_1} and H_{l_2} separately:

$$M_{l_1}^{'} = \sigma(H_{l_1} \cdot W_{l_1}^{\top} + b_{l_1}^{'}) \,,$$

$$M_{l_1}^{*} = \sigma(H_{l_2} \cdot W_{l_1}^{\top} + b_{l_1}^{'}) \,,$$

$$M_{l_2}^{'} = \sigma(H_{l_2} \cdot W_{l_2}^{\top} + b_{l_2}^{'}) \,,$$

$$M_{l_2}^{*} = \sigma(H_{l_1} \cdot W_{l_2}^{\top} + b_{l_2}^{'}) \,,$$

where $b_{l_1}^{'}$, $b_{l_2}^{'}$ are the bias vectors. $M_{l_1}^{'}$ and $M_{l_2}^{'}$ are the monolingual reconstructions of M_{l_1} and M_{l_2} from the common space, and $M_{l_1}^{*}$ and $M_{l_2}^{*}$ are cross-lingual reconstructions. $W_{l_1}^{\top}$ and $W_{l_2}^{\top}$ are the transposes of W_{l_1} and W_{l_2} respectively.

To learn the common semantic representations, we minimize the distance between the aligned word vectors as well as the loss of monolingual and cross-lingual reconstruction:

$$O_W = \sum_{\{l_i, l_j\} \in A} L(M_{l_i}^{'}, M_{l_i}) + L(M_{l_i}^{*}, M_{l_i})$$

$$+ L(M_{l_j}^{'}, M_{l_j}) + L(M_{l_j}^{*}, M_{l_j}) + L(H_{l_i}, H_{l_j}) \,,$$

where l denotes any language that we want to project into the common semantic space, A denotes all bilingual dictionaries, and L denotes a

similarity metric. In our work, we use cosine similarity as the similarity metric.

3.3 Neighborhood-Consistent CorrNet

CorrNet can project multiple monolingual word embeddings into a common semantic space using bilingual word alignment. However, the same concepts may have different semantic bias in various languages. For example, the top five nearest words of the concept "*China*" are: (*Japan, India, Taiwan, Chinese, Asia*) in English, (*Cosco, Shenzhen, Australian, Shanghai, manufacturing*) in Danish, and (*Beijing, Korea, Japan, aluminum, copper*) in Italian respectively. In order to ensure the consistency of the neighborhoods within the common semantic space and make the cross-lingual mapping locally smooth, we propose to augment monolingual word representation with its top-N nearest neighboring words from the original monolingual semantic space.[3]

Given the monolingual embeddings of the bilingual aligned words for two languages l_1 and l_2, M_{l_1} and M_{l_2}, for each word, we extract the top-N nearest neighbors and construct the neighborhood clusters. Each cluster $t_l = \{w_1, w_2, ..., w_{|t_l|}\}$ in language l is represented by

$$c_{t_l} = \frac{1}{|t_l|} \sum_{w \in t_l} E_w \,,$$

where E_w denotes the monolingual word embedding for w.

We obtain all the neighborhood cluster vector representations C_{l_1}, C_{l_2} for l_1 and l_2. We incorporate the neighborhood cluster information into the common semantic space when projecting monolingual embeddings:

$$\begin{aligned} H_{l_1} &= \sigma(M_{l_1} \cdot W_{l_1} + C_{l_1} \cdot U_{l_1} + b_{l_1}), \\ H_{l_2} &= \sigma(M_{l_2} \cdot W_{l_2} + C_{l_2} \cdot U_{l_2} + b_{l_2}), \end{aligned} \quad (1)$$

Besides the monolingual and cross-lingual reconstructions for M_{l_1} and M_{l_2} in CorrNets, we also add monolingual and cross-lingual reconstructions for the neighborhood clusters:

$$C'_{l_1} = \sigma(H_{l_1} \cdot U_{l_1}^\top + b^*_{l_1}) \,,$$

$$C^*_{l_1} = \sigma(H_{l_2} \cdot U_{l_1}^\top + b^*_{l_1}) \,,$$

$$C'_{l_2} = \sigma(H_{l_2} \cdot U_{l_2}^\top + b^*_{l_2}) \,,$$

$$C^*_{l_2} = \sigma(H_{l_1} \cdot U_{l_2}^\top + b^*_{l_2}) \,,$$

In addition to optimizing the loss functions described in the Section 3.2, we further optimize the monolingual and cross-lingual reconstruction for neighborhood clusters:

$$\begin{aligned} O_N = \sum_{\{l_i, l_j\} \in A} & L(C'_{l_i}, C_{l_i}) + L(C^*_{l_i}, C_{l_i}) \\ & + L(C'_{l_j}, C_{l_j}) + L(C^*_{l_j}, C_{l_j}) \,, \end{aligned}$$

3.4 Character-Level Word Alignment

Bilingual word alignment is not always enough or available to induce a common semantic space, especially for low-resource languages. Although the words that refer to the same concept are not exactly the same in multiple languages, they often share a set of similar characters, especially in related languages written in the same script. For example, the same entity is spelled slightly differently in three languages: *Semsettin Gunaltay* in English, *Şemsettin Günaltay* in Turkish, and *Semsetin Ganoltey* in Somali. Beyond word-level alignment, we introduce character-level alignment by composing word representations from its compositional characters using convolutional neural networks (CNN). For each language, we adopt a language-independent CNN to generate character-level word representation.

Character Lookup Embeddings Let S_l be the character set for language l and $E_{S_l} \in \mathbb{R}^{|S_l| \times d}$ be the character lookup embeddings, where d is the dimensionality of each character embedding. Here, we use a simple yet effective method to induce character embeddings from word embeddings[4]. For each character c, we initialize its embedding by averaging the embeddings of all words which contain the character. The character embeddings will be further tuned by the model.

Character-Level CNN (Kim et al., 2016) The input layer is a sequence of characters of length k for each word. Each character is represented by a d-dimensional lookup embedding. Thus each input sequence is represented as a feature map of dimensionality $d \times k$.

We use the convolution layer to learn the representation for each sliding n-gram characters. We make p_i as the concatenated embeddings of n continuous columns from the input matrix, where n

[3] We set $N = 10$ in our experiments since it performed best on the intrinsic evaluation among {2, 5, 10, 20, 50}.

[4] This approach is proved to be better than random initialization of character embeddings.

is the filter width. We then apply the convolution weights $W \in \mathbb{R}^{d \times nd}$ to p_i with a biased vector $b \in \mathbb{R}^d$, i.e., $p_i^{'} = \tanh(W \cdot p_i + b)$. All n-gram representations $p_i^{'}$ are used to generate the word representation y by max-pooling.

In our experiments, we apply multiple filters with various widths to obtain the representation for word w_i^l. The final character-level word representation $\hat{w}_i^l$ is the concatenation of all word representations with varying filter widths.

Cross-lingual Mapping Given the bilingual aligned word pairs, we directly minimize the distance of the character-level word representations in the common semantic space by:

$$O_{char} = \sum_{\{l_i, l_j\} \in A} L(\hat{W}_{l_i}^{char}, \hat{W}_{l_j}^{char})$$

The final word representation of w_i^l in the common semantic space is the concatenation of character-level word presentation $\hat{w}_i^l$ and projected word representation h_i^l.

3.5 Linguistic Property Alignment

Linguists have made great efforts at building linguistic property knowledge bases for thousands of languages in the world. These knowledge bases include a large number of topological properties (phonological, lexical and grammatical) which we will use to build a high-level alignment between words across languages. We exploit the following resources:

- **CLDR** (Unicode Common Locale Data Repository)[5] which includes multilingual gazetteers for months, weekdays, cardinal and ordinal numbers;
- **Wiktionary**[6] which is a multilingual, web-based collaborative project to create an English content dictionary, includes word and prefix/suffix dictionaries for 1,247 languages;
- **Panlex**[7] database which contains 1.1 billion pairwise translations among 21 million expressions in about 10,000 language varieties.

We mainly exploit two types of linguistic properties to extract word clusters. The first type is language-independent closed word classes, such as colors, weekdays, and months. Table 1 shows some examples of the word clusters we automatically extracted from CLDR and Wiktionary for English. The second type of word clusters is generated based on morphological information, including affixes that indicate various linguistic properties. These properties tend to be consistent across many languages. For example, "-like" is a suffix denoting "similar to" in English, while in Danish "-agtig" performs the same function. Wiktionary and Panlex include the affix alignments between English and any other languages. We filtered out the many-to-many affix alignments and obtained hundreds of alignments between each language and English. For each affix, we derive a set of word pairs (*basic word*, *extended word with affix*) by first selecting all the word pairs where *basic word* + *affix* = *extended word*, then ranking all word pairs based on the cosine similarity of their monolingual word embedding. Finally we select the top ranked 20 word pairs to form the cluster for each affix.

Class Name	Words / Word Pairs
Colors	white, yellow, red, blue, green ...
Weekdays	monday, tuesday, friday, sunday ...
Months	january, february, march, april ...
numbers	one, two, three, four, five ...
pronouns	i, me, you, he, she, her, they ...
prepositions	of, in, on, for, from, about ...
conjunctions	but, and, so, or, when, while ...
clothes	hat, shirt, pants, skirt, socks ...
-like	(god, godlike), (bird, birdlike) ...
-able	(accept, acceptable), (adopt, adoptable) ...
micro-	(gram, microgram), (chip, microchip) ...
auto-	(maker, automaker), (gas, autogas) ...

Table 1: Examples of closed word classes and linguistic properties based clusters for English

We extract a set of word clusters from each language, and align the clusters based on their functions defined in CLDR, Wiktionary and Panlex. For each language l, each cluster $r_i^l \in R^l$ contains a set of words or word-pairs sharing the same function. We use the average operation to obtain an overall vector representation for each cluster M_l^R.[8] Then, we project the cluster-level vectors into the shared semantic space and minimize the distance between them:

$$H_{l_i}^R = \sigma(M_{l_i}^R \cdot W_{l_i} + b_{l_i}^R)\,,$$

$$H_{l_j}^R = \sigma(M_{l_j}^R \cdot W_{l_j} + b_{l_j}^R)\,,$$

[5] http://cldr.unicode.org/index/charts

[6] https://en.wiktionary.org

[7] http://panlex.org/

[8] For each word pair, we use the vector of the extend word minus the vector of the basic word as the vector representation of the word pair.

Parameter Name	Value
Monolingual Word Embedding Size	512
Multilingual Word Embedding Size	512
# of Filters in Convolution Layer	20
Filter Widths	1, 2, 3
Batch Size	500
Initial Learning Rate	0.5
Optimizer	Adadelta

Table 2: Hyper-parameters.

$$O_R = \sum_{\{l_i,l_j\}\in A} L(H_{l_i}^R, H_{l_j}^R)\,,$$

where W is the same as the W used in Section 3.3 for each language. We finally optimize the sum of the losses by finding the parameters $\theta = \{W_l, b_l, b_l', U_l, b_l^*, \text{CNN}_l, b_l^R\}$, where l denotes a specific language:

$$O_\theta = O_W + O_N + O_{char} + O_R$$

4 Experiments

4.1 Experiment Setup

Previous work (Ammar et al., 2016b; Duong et al., 2017) evaluated multilingual word embeddings on a series of intrinsic (e.g., monolingual and cross-lingual word similarity, word translation) and extrinsic (e.g., multilingual document classification, multilingual dependency parsing) evaluation tasks. In order to evaluate the quality of the multilingual embeddings, we use QVEC (Tsvetkov et al., 2015) tasks (details will be described in Section 4.2) as the intrinsic evaluation platform. In addition, to demonstrate the effectiveness of our common semantic space for knowledge transfer, especially for low-resource scenarios, we adopt the low-resource language name tagging task for extrinsic evaluation.

For fair comparison with state-of-the-art methods on building multi-lingual embeddings (Ammar et al., 2016b; Duong et al., 2017), we use the same monolingual data and bilingual dictionaries as in their work. We build multilingual word embeddings for 3 languages (*English, Italian, Danish*) and 12 languages (*Bulgarian, Czech, Danish, German, Greek, English, Spanish, Finnish, French, Hungarian, Italian, Swedish*) respectively. The monolingual data for each language is the combination of the Leipzig Corpora Collection[9] and Europarl.[10] The bilingual dictionaries are the same as those used in Ammar et al. (2016b).[11]

For each task, we evaluate the performance of our common semantic space in comparison with previously published multilingual word embeddings (MultiCluster, MultiCCA, MultiSkip, and MultiCross)[12]. MultiCluster (Ammar et al., 2016b) groups multilingual words into clusters based on bilingual dictionaries and forces all the words from various languages within one cluster share the same embedding. MultiCCA (Ammar et al., 2016b; Faruqui and Dyer, 2014) uses CCA to estimate linear projections for each pair of languages. MultiSkip is an extension of the multilingual skip-gram model (Luong et al., 2015), which requires parallel data. MultiCross is an approach to unify bilingual word embeddings into a shared semantic space using post hoc linear transformations (Duong et al., 2017).

Table 2 lists the hyper-parameters used in the experiments.

4.2 Intrinsic Evaluation: QVEC

In order to evaluate the quality of multilingual embeddings, we adopt QVEC (Tsvetkov et al., 2015) as the intrinsic evaluation measure. It evaluates the quality of word embeddings based on the alignment of distributional word vectors to linguistic feature vectors extracted from manually crafted lexical resources, e.g., SemCor (Miller et al., 1993). For each word, each dimension of its linguistic feature vector defines the probability of that word belongs to a supersense (e.g., NN.FOOD) which is summarized from WordNet (Fellbaum, 1998).

QVEC is computed as

$$\text{QVEC} = \max_{\sum_j a_{ij}\le 1} \sum_{i=1}^{D}\sum_{j=1}^{P} r(x_i, s_j) \times a_{ij},$$

where $x \in \mathbb{R}^{D\times 1}$ denotes a distributional word vector and $s \in \mathbb{R}^{P\times 1}$ denotes a linguistic word vector. D and P denote the sizes of vectors respectively. $a_{ij} = 1$ iff x_i is aligned to s_j, otherwise $a_{ij} = 0$. $r(x_i, s_j)$ is the Pearson's correlation between x_i and s_j. QVEC-CCA (Ammar et al., 2016b) is extended from QVEC by using CCA to measure the correlation between the distributional matrix and the linguistic vector matrix, instead of cumulative dimension-wise correlation.

[9] http://wortschatz.uni-leipzig.de/en/download/
[10] http://www.statmt.org/europarl/index.html
[11] http://128.2.220.95/multilingual/data/

[12] For fair comparison, we use the development sets of the intrinsic evaluation tasks in Ammar et al. (2016b) to select the best model.

		3 Languages				12 Languages			
		Monolingual		Multilingual		Monolingual		Multilingual	
		QVEC	QVEC-CCA	QVEC	QVEC-CCA	QVEC	QVEC-CCA	QVEC	QVEC-CCA
MultiCluster		10.8	63.6	9.1	**45.8**	10.4	62.7	9.3	44.5
MultiCCA		10.8	63.8	8.5	43.9	10.8	63.9	8.5	43.7
MultiSkip		7.8	57.3	7.3	36.2	8.4	59.1	7.2	36.5
MultiCross		-	-	-	-	11.9	46.4	8.6	31.0
CorrNet	W	14.8	63.6	11.3	43.4	14.7	63.8	13.2	43.9
	W+N	15.8	64.2	**13.1**	43.9	15.8	64.6	14.0	44.8
	W+N+C	15.3	66.2	12.2	44.5	16.0	66.6	14.0	44.7
	W+N+L	**16.2**	66.1	**13.1**	44.8	16.1	64.7	13.8	44.9
	W+N+C+L	**16.2**	**67.3**	12.4	45.4	**16.3**	**66.7**	**14.1**	**45.2**

Table 3: QVEC and QVEC-CCA scores. W: word alignment. N: neighbor based clustering and alignment. C: character based clustering and alignment. L: linguistic property based clustering and alignment.

As shown in Table 3, our approaches outperform previous approaches in almost all cases[13]. Specifically, by augmenting word representation with neighboring words in the common semantic space as in Eq. (1), the performance for monolingual and multilingual QVEC and QVEC-CCA tasks is consistently improved. In addition, by aligning character-level compositional representations and linguistic property based clusters in the shared semantic space, the monolingual and multilingual representation quality is further improved.

4.3 Impact of Bilingual Dictionary Size

In order to show the impact of the size of bilingual lexicons, we use three languages as a case study, and gradually reduce the size of the lexicons for each pair of languages from 40,000 to 10,000 and further to 2,000, 1,000, 500 and 250. For following experiments, we use MultiCluster and MultiCCA as baselines [14]. Table 4 shows the results. We observe that both MultiCCA and CorrNet approaches are sensitive to the size of the bilingual lexicons. Our approach on the other hand can maintain high performance, even when the size of bilingual lexicons is reduced to 250. The performances of MultiCluster based on various sizes of bilingual dictionary are close because it jointly trains the embedding of multiple languages from scratch and by default takes advantage of identical strings among all the languages.

[13]We conduct paired t-test between CorrNet W+N+Ch+L and all the other models on 10 randomly sampled subsets. The differences are all statistically significant while all p-values are less than 0.05

[14]MultiSkip requires parallel corpora to train cross-lingual embeddings while the original implementation of MultiCross is not public.

4.4 Low-Resource Name Tagging

We evaluate the quality of multilingual embeddings on a downstream task by using the embeddings as input features. Here, we use low-resource language name tagging as a target task, which aims to automatically identify and named entities from text and classify them into certain types, including Person (PER), Location (LOC), Organization (ORG), and Geo-Political Entities (GPE). We experiment with two sets of languages. The first set *Amh+Tig* consists of Amharic and Tigrinya. Both languages share the same Ge'ez script and descend from the proto-Semitic language family. The other set *Eng+Uig+Tur* consists of one high-resource language (English), one medium-resource language (Turkish) and one low-resource language (Uighur). It also consists of two distinct language scripts: English and Turkish use Latin script while Uighur uses Arabic script.

We use an LSTM-CRF architecture (Huang et al., 2015; Lample et al., 2016; Ma and Hovy, 2016) for name tagging. It takes only word embedding as input and predict a tag for each word. Table 5 shows the statistics of training, development, and test sets for each language released by Linguistic Data Consortium (LDC).[15] For each language pair, we combine the bilingual aligned words extracted from Wiktionary and monolingual dictionaries based on identical strings.[16] We evaluate the quality from several aspects:

[15]The annotations are from: Amh (LDC2016E87), Tig (LDC2017E27), Uig (LDC2016E70), Tur (LDC2014E115), Eng (Tjong Kim Sang and De Meulder, 2003). We combined these corpora with Wikipedia dump to train word embeddings with Word2Vec toolkit (Mikolov et al., 2013a).

[16]We extracted 23,781 pairs of words for Amh and Tig, 16,868 pairs for Eng and Tur, 3,353 pairs for Eng and Uig, and 3,563 pairs for Tur and Uig.

		QVEC		QVEC-CCA	
		Monolingual	Multilingual	Monolingual	Multilingual
40,000	multiCCA	10.8	8.5	63.8	43.9
	multiCluster	10.8	9.1	63.6	**45.8**
	CorrNet W	14.8	11.3	63.6	43.4
	CorrNet W+N+C+L	**16.2**	**12.4**	**67.3**	45.4
10,000	multiCCA	9.8	6.5	63.6	42.3
	multiCluster	10.6	9.5	62.4	44.7
	CorrNet W	14.8	11.3	63.4	43.0
	CorrNet W+N+C+L	**15.7**	**12.4**	**68.0**	**45.1**
2,000	multiCCA	9.9	6.2	63.6	40.9
	multiCluster	10.5	9.3	62.5	**44.8**
	CorrNet W	**14.5**	7.1	62.0	39.2
	CorrNet W+N+C+L	**14.5**	**11.4**	**68.0**	**44.8**
1,000	multiCCA	12.3	6.9	63.5	38.2
	multiCluster	10.5	9.3	62.5	**44.8**
	CorrNet W	**13.7**	9.4	63.0	40.0
	CorrNet W+N+C+L	13.6	**10.5**	**66.4**	43.0
500	multiCCA	12.3	5.5	63.5	36.0
	multiCluster	10.5	9.3	62.6	**44.7**
	CorrNet W	13.3	9.1	62.8	39.4
	CorrNet W+N+C+L	**13.4**	**9.5**	**66.2**	42.7
250	multiCCA	12.3	5.3	63.5	35.0
	multiCluser	10.5	9.2	62.7	**44.9**
	CorrNet W	13.8	9.3	62.5	39.3
	CorrNet W+N+C+L	**13.9**	**9.8**	**65.9**	42.2

Table 4: Results using bilingual lexicons with varying sizes (40,000, 10,000, 2,000, 1,000, 500, 250) and three languages. CorrNet W+N+C+L is the proposed approach with all the cluster types.

	Amh	Tig	Uig	Tur	Eng
Train	1,506	1,585	1,500	1,500	14,029
Dev	167	176	190	378	3,250
Test	711	440	476	470	3,453

Table 5: # of Sentences for name tagging

		Multilingual			
Train & Test	Mono-lingual	Multi-CCA	Multi-Cluster	CorrNet W	CorrNet W+N+C+L
Amh	52.0	50.6	53.4	52.4	**55.8**
Tig	78.2	78.4	76.4	77.9	**78.5**
Uig	63.3	59.6	60.1	61.9	**65.2**
Tur	62.9	47.7	54.0	59.3	**64.9**

Table 6: Comparison on Monolingual Embedding Quality: name tagging performance (F-score, %) using monolingual embedding and multilingual embeddings.

Monolingual embedding quality evaluation Table 6 shows the name tagging performance for each language using the original monolingual embeddings and multilingual embeddings. For all languages, the multilingual embeddings learned from our approach significantly outperform those learned from MultiCCA and MultiCluster, which shows the effectiveness of our approach. More importantly, the multilingual embeddings learned from our approach also outperform original monolingual embeddings, which demonstrates that by projecting multiple languages into one common space, the monolingual embedding quality can be further improved.

Cross-lingual direct transfer In this setting, we train a name tagger on one or two languages using multilingual embeddings and test it on a new language without any annotated data. Table 7 shows the performance. For most testing languages, our approach achieves better performance than MultiCCA and MultiCluster. The closer that the languages are, such as Amharic and Tigrinya, the better performance is achieved.

Cross-lingual mutual enhancement We finally show the improvement by adding more cross-lingual annotated data and also using multilingual embeddings in Table 8. The multilingual embeddings learned by our approach consistently outperforms MultiCCA and MultiCluster. Particularly, when there are not enough annotated examples, the performance could be improved by incorporating annotated examples from other languages. This is evident for Amharic, Tigrinya and Uighur. For Turkish, we notice that a larger extra anno-

Train	Test	Multi-CCA	Multi-Cluster	CorrNet W	CorrNet W+N+C+L
Amh	Tig	15.5	29.7	28.3	**33.7**
Tig	Amh	11.1	**24.7**	12.8	23.3
Eng	Uig	4.8	9.1	13.3	**15.5**
Tur	Uig	0.4	11.4	19.8	**25.0**
Eng+Tur	Uig	8.3	10.5	17.3	**23.3**
Eng	Tur	17.6	21.4	18.3	**22.4**
Uig	Tur	6.9	12.8	**13.2**	10.7
Eng+Uig	Tur	20.4	23.3	14.5	**27.0**

Table 7: Comparison on Cross-lingual Direct Transfer: name tagging performance (F-score, %) when the tagger was trained on 1-2 source languages and tested on a target language.

Train	Test	Multi-CCA	Multi-Cluster	CorrNet W	CorrNet W+N+C+L
Tig+Amh	Amh	52.9	54.7	52.1	**56.5**
Amh+Tig	Tig	78.0	76.9	78.1	**78.7**
Eng+Uig	Uig	64.8	62.2	65.1	**67.7**
Tur+Uig	Uig	63.6	58.9	63.6	**65.8**
Eng+Tur+Uig	Uig	65.8	64.8	64.6	**68.5**
Eng+Tur	Tur	50.3	56.1	59.3	**65.5**
Uig+Tur	Tur	51.4	52.7	57.8	**62.7**
Eng+Uig+Tur	Tur	48.1	54.3	56.6	**61.5**

Table 8: Comparison on Cross-lingual Mutual Enhancement: name tagging performance (F-score, %) when the training set for the tagger was enhanced by annotated examples in other languages.

tated set from other languages (e.g., Uig+Tur or Eng+Uig+Tur) doesn't necessarily result in improvement. This is partially due to the use of Arabic script in Uighur, which differs from Turkish and English. Thus we suggest to project closely related languages using the same script into the common semantic space.

We take Turkish name tagging as a case study to show the benefit of the common semantic space with extra English annotations. The monolingual model failed to identify *Belgrad´da* as a geopolitical entity (GPE) because it doesn't occur in Turkish training data. However, by adding English annotations, the tagger successfully tags it as a GPE since it's semantically close to *Belgrade* in the common semantic space according to their character level compositional embeddings and *Belgrade* is frequently tagged as GPE in English annotations. In another example, using Turkish annotations only, *Kraliyet Donanması´na* is mistakenly tagged as a GPE since it's following *da* and all entity mentions following *da* in Turkish annotations are annotated as GPE. After adding English annotations into training, it is correctly tagged as an ORG because *da* is well aligned with *in* in the common semantic space according to the linguistic property alignment between Turkish and English, and many entity mentions following *in* are annotated as ORG in English annotations.

5 Conclusions and Future Work

We construct a common semantic space for multiple languages based on a cluster-consistent correlational neural network. It combines word-level alignment and multi-level cluster alignment, including neighbor based clusters, character-level compositional word representations, and linguistic property based clusters induced from the readily available language-universal linguistic knowledge bases. Our approach achieved significantly higher performance than state-of-the-art multilingual embedding learning methods through both intrinsic and extrinsic evaluations. In the future, we will further extend our approach to multi-lingual multimedia common semantic space construction.

Acknowledgments

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract # FA8650-17-C-9116, and U.S. DARPA LORELEI Program # HR0011-15-C-0115. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. Kyunghyun Cho also thanks support by eBay, TenCent, NVIDIA and CIFAR.

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2016a. Many languages, one parser. *arXiv preprint arXiv:1602.01595*.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016b. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Sarath AP Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C

Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Proceedings of NIPS*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of ACL*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297*.

Marco Baroni, Angeliki Lazaridou, and Georgiana Dinu. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning.

Hailong Cao, Tiejun Zhao, Shu Zhang, and Yao Meng. 2016. A distribution-based model to learn bilingual word embeddings. In *Proceedings of COLING*.

Sarath Chandar, Mitesh M Khapra, Hugo Larochelle, and Balaraman Ravindran. 2016. Correlational neural networks. *Neural computation*.

Leon Cheung, Thamme Gowda, Ulf Hermjakob, Nelson Liu, Jonathan May, Alexandra Mayn, Nima Pourdamghani, Michael Pust, Kevin Knight, Nikolaos Malandrakis, et al. Elisa system description for lorehlt 2017.

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

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. *arXiv preprint arXiv:1606.09403*.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2017. Multilingual training of crosslingual word embeddings. In *Proceedings of EMNLP*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of EMNLP*.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Xiaocheng Feng, Lifu Huang, Bing Qin, Ying Lin, Heng Ji, and Ting Liu. 2017. Multi-level cross-lingual attentive neural architecture for low resource name tagging. *Tsinghua Science and Technology*, 22(6):633–645.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of ICML*.

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

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributional semantics. In *Proceedings of ACL*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

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

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of HLT-NAACL*.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of HLT-NAACL*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Pranava Swaroop Madhyastha and Cristina España-Bonet. 2017. Learning bilingual projections of embeddings for vocabulary expansion in machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*.

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*.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of HLT*.

Janarthanan Rajendran, Mitesh M Khapra, Sarath Chandar, and Balaraman Ravindran. 2015. Bridge correlational neural networks for multilingual multimodal representation learning. *arXiv preprint arXiv:1510.03519*.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. *arXiv preprint arXiv:1602.07572*.

Holger Schwenk, Ke Tran, Orhan Firat, and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. *arXiv preprint arXiv:1704.04154*.

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

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of NAACL-HLT*.

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

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of EMNLP*.

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

Ivan Vulic and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of ACL*.

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

Boliang Zhang, Ying Lin, Xiaoman Pan, Di Lu, Jonathan May, Kevin Knight, and Heng Ji. 2018. Elisa-edl: A cross-lingual entity extraction, linking and localization system. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 41–45.

Boliang Zhang, Xiaoman Pan, Ying Lin, Tongtao Zhang, Kevin Blissett, Samia Kazemi, Spencer Whitehead, Lifu Huang, and Heng Ji. 2017a. Rpi blender tac-kbp2017 13 languages edl system.

Boliang Zhang, Xiaoman Pan, Ying Lin, Tongtao Zhang, Kevin Blissett, Samia Kazemi, Spencer Whitehead, Lifu Huang, and Heng Ji. 2017b. Rpi blender tac-kbp2017 13 languages edl system. In *TAC*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017c. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of ACL*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017d. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of EMNLP*.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi S Jaakkola. 2016. Ten pairs to tag-multilingual pos tagging via coarse mapping between embeddings. In *Proceedings of HLT-NAACL*.

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