Improving Vector Space Word Representations Via Kernel Canonical Correlation Analysis

XUEFENG BAI, HAILONG CAO, and TIEJUN ZHAO, Harbin Institute of Technology

Cross-lingual word embeddings are representations for vocabularies of two or more languages in one common continuous vector space and are widely used in various NLP tasks. A simple yet efficient way to generate cross-lingual word embeddings is using canonical correlation analysis (CCA). However, CCA works with the assumption that the vector representations of similar words in different languages are related by a linear relationship. This assumption does not always hold true, especially for substantially different languages. We therefore propose to use kernel canonical correlation analysis (KCCA) to capture non-linear relationships between word embeddings of two languages. By extensively evaluating the resulting word embeddings on three tasks (word similarity, cross-lingual dictionary induction, cross-lingual document classification) across five language pairs, we show that our approach produces essentially better semantic vectors than CCA-based method, especially for substantially different languages.

CCS Concepts: • **Computing methodologies → Machine translation**;

Additional Key Words and Phrases: Cross-lingual word representation, kernel canonical correlation analysis (KCCA), word embedding evaluation

ACM Reference format:
XueFeng Bai, HaiLong Cao, and TieJun Zhao. 2017. Improving Vector Space Word Representations Via Kernel Canonical Correlation Analysis. 1, 1, Article 1 (June 2017), 14 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Monolingual word embeddings have made great achievements in many NLP tasks including sentiment analysis (Socher et al. 2013) , dependency parsing (Dyer et al. 2015, Guo et al. 2015) . As a natural extension of monolingual word embeddings, cross-lingual word embeddings can not merely improve the performance on monolingual NLP tasks, but facilitate some cross-lingual tasks as well, such as machine translation (Uszkoreit et al. 2010, Mikolov et al. 2013b, Zhang et al. 2014), cross-lingual document classification (Wan 2009, Klementiev et al. 2012, Shi et al. 2015,Upadhyay et al. 2016), cross-lingual dependency parsing (McDonald et al. 2013, Gouws et al. 2015).

Several models for inducing cross-lingual word embeddings have been proposed recently. From an optimization perspective, these models can be roughly classified into two categories. In the

This work is supported by the National Science Foundation, under grant CNS-0435060, grant CCR-0325197 and grant EN-CS-0329609.

Author’s addresses: X. Bai, H. Cao, and T. Zhao, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China; emails: {xfbai, hailong, tjzhao }@mtlab.hit.edu.cn.

XXXX-XXXX/2017/6-ART1 $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

first category (Mikolov et al. 2013b, Faruqui and Dyer 2014), monolingual corpora of each language are used to train word representations, which are then transformed into cross-lingual word representations as a separate step. In the second category (Hermann and Blunsom 2014, Shi et al. 2015, Gouws et al. 2015, Luong et al. 2015), two monolingual models are jointly trained, with the cross-lingual objective enforced as constraints. Compared to the second category, the first type is often computationally-efficient and easy to scale to large datasets. BiCCA proposed by Faruqui and Dyer is a typical model in the first category. The idea of this model is that assuming there is often a strong linear relationship between two translationally equivalent words, this model captures these linear factors using canonical correlation analysis (CCA). However, we have found a lot of non-linear relationships which can not be captured by CCA, especially among substantially different languages. Hence we propose a more powerful model to capture non-linear relationships using kernel canonical correlation analysis (KCCA) (Akaho 2006).

The overall structure of the article is as follows. A brief introduction to Faruqui and Dyer's model is firstly given as background in Section 2. In Section 3, we present illustrations of our proposed model, together with an explanation of the training process. Section 4 describes the evaluation tasks — our intrinsic evaluation assesses the quality of the vectors on monolingual (word similarity for English) and cross-lingual (cross-lingual dictionary induction) tasks, while our extrinsic evaluation (cross-lingual document classification) assesses the ability of cross-lingually trained vectors to facilitate model transfer across languages. In Section 5, experimental methodology — the corpora used, the setting of our model, and the results of the evaluation tasks are present. In Section 6, a systematic analysis of the resulting word embeddings, errors and the proposed model is given, with examples from the experimental data. We conclude and offer possible directions for future research in Section 7.

Our contributions are the following:

- We propose a KCCA-based method to capture non-linear relationships between two languages.
- We extensively evaluate embeddings produced by our model on both extrinsic and intrinsic tasks across five language pairs, and show our method produces essentially higher-quality vectors than original linear model.
- We qualitatively analyze the resulting word embeddings, and discover that there are a lot nonlinear relationship among substantially different languages, while little among closely related languages. We believe this discovery will be helpful for further research.

2 BACKGROUND: BILINGUAL CORRELATION BASED EMBEDDINGS (BICCA)

In the following, we briefly introduce the BiCCA model (Faruqui and Dyer 2014), which transforms monolingual trained word embeddings into bilingual word embeddings using CCA. BiCCA generates new word representations for each language which incorporates cross-lingual knowledge from two source languages.

The setup of BiCCA is that assuming monolingual word embeddings for two languages, denoted by $\Sigma \in \mathbb{R}^{n_1 \times d_1}$, $\Omega \in \mathbb{R}^{n_2 \times d_2}$, together with a bilingual lexicon are provided initially. BiCCA then generates bilingual word embeddings by projecting original monolingual vector matrices Σ, Ω onto a shared vector space. The embeddings generated are expected to hold the following properties in shared vector space: (i) semantically similar words in the same language are nearby. (ii) translationally equivalent words in different languages are nearby.

Let $\Sigma \in \mathbb{R}^{n_1 \times d_1}$ and $\Omega \in \mathbb{R}^{n_2 \times d_2}$ be word vector matrices of two languages vocabularies, n_1, n_2 represent the size of vocabularies, d_1, d_2 denote the dimensionality of word vectors. The training

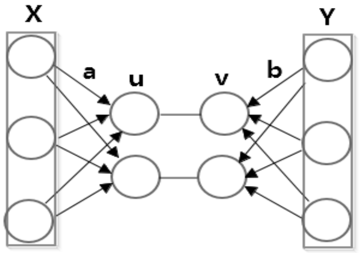

Fig. 1. CCA seeks a pair of linear transformations $a \in \mathbb{R}^{n_x}$, $b \in \mathbb{R}^{n_y}$ in order to maximize the correlation coefficient $\rho = corr(u, v)$, in which $u = a^T X, v = B^T Y$

matrices Σ', Ω' are firstly constructed as $\Sigma' \subset \Sigma$, $\Omega' \subset \Omega$, such that $|\Sigma'| = |\Omega'|$ and the corresponding words (Σ'_i, Ω'_i) in the matrices are translations of each other. The projection matrix P_Σ, P_Ω is then computed as:

$$P_\Sigma, P_\Omega = CCA(\Sigma', \Omega') \tag{1}$$

where $P_\Sigma \in \mathbb{R}^{d_1 \times k}$, $P_\Omega \in \mathbb{R}^{d_1 \times k}$, $k \leq min\{d_1, d_2\}$, CCA is introduced latter in the following part of this section. Finally, original monolingual vector matrices are projected as:

$$\Sigma^* = \Sigma P_\Sigma, \Omega^* = \Omega P_\Omega \tag{2}$$

where $\Sigma^* \in \mathbb{R}^{n_1 \times k}, \Omega^* \in \mathbb{R}^{n_2 \times k}$ are the word vectors that have been "enriched" using bilingual knowledge.

Canonical Correlation Analysis A popular method for multi-representation learning is canonical correlation analysis (Hotelling 1936), which is a method of correlating linear relationships between two multidimensional variables. It finds two projection vectors, one for each variable, which are optimal with respect to correlations. The dimension of these new projected vectors is equal to or less than the smaller dimension of the two variables.

Given a pair of multi-variates $X \in \mathbb{R}^{n_x}$, $Y \in \mathbb{R}^{n_y}$, CCA[1] finds a pair of projection vectors $a \in \mathbb{R}^{n_x}$, $b \in \mathbb{R}^{n_y}$, let $u = a^T X, v = B^T Y$, such that the correlation coefficient $\rho = corr(u, v)$ is maximized. (Fig. 1)

Let $\Sigma_{xx} = cov(X, X)$, $\Sigma_{yy} = cov(Y, Y)$ and $\Sigma_{xy} = cov(X, Y)$, the object to maximize is:

$$\rho(u, v) = \frac{cov(u, v)}{\sqrt{Var(u)}\sqrt{Var(v)}} = \frac{a^T \Sigma_{xy} b}{\sqrt{a^T \Sigma_{xx} a}\sqrt{b^T \Sigma_{yy} b}} \tag{3}$$

Observing that Equation (3) is not affected by the rescaling of a and b either together or independently, the CCA optimization problem is equivalent to maximizing the numerator subject to:

$$Var[u] = Var[v] = 1 \tag{4}$$

The corresponding Lagrangian to this optimization problem is:

$$L = a^T \Sigma_{xy} b - \frac{\lambda}{2}(a^T \Sigma_{xx} a - 1) - \frac{\theta}{2}(b^T \Sigma_{yy} b - 1) \tag{5}$$

[1] For more details of CCA, please refer to Relations Between Two Sets of Variates (Hotelling 1936)

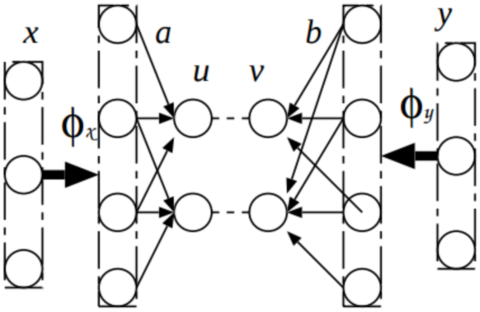

Fig. 2. KCCA seeks two projection vectors $a \in H_x$, $b \in H_y$ such that the inner products $u = a^T \Phi_x(X), v = b^T \Phi_y(Y)$ maximize the correlation coefficient $\rho = corr(u, v)$.

By solving Equation (5), a and b can be found by an eigenvector corresponding to the maximal eigenvalues of a generalized eigenvalue problem:

$$\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}a = \lambda^2 a \tag{6}$$

If more dimension is needed, we can take eigenvectors corresponding to other maximal eigenvalues.

3 OUR APPROACH

BiCCA introduced in §(2) has been proved effective to generate cross-lingual word embeddings, however, as we will show later, there are a lot of non-linear relationships between words in the source and target languages which can not be learned by BiCCA. Hence, we propose a improved model based on kernel canonical correlation analysis (KCCA) to capture the non-linear relationship.

3.1 Kernel Canonical Correlation Analysis

In this subsection, we introduce KCCA briefly.[2] The kernel version of CCA offers an alternate solution to learn non-linear factors by first projecting the data X, Y onto Hilbert spaces H_x, H_y via kernel function Φ_x, Φ_y:

$$\Phi : X = (X_1 X_2 \cdots X_n) \longrightarrow \Phi(X) = (\phi_1(X), \phi_2(X) \cdots \phi_N(X)), \text{n<N}.$$

which could be a non-linear transformation, and then performing CCA in the new feature space.

First, X and Y are transformed into Hilbert space, $\Phi_x(X) \in H_x, \Phi_y(Y) \in H_y$. By taking inner products with projection vectors $a \in H_x, b \in H_y$, we find two features:

$$u = a^T \Phi_x(X) \tag{7}$$

$$v = b^T \Phi_y(Y) \tag{8}$$

which maximize the correlation coefficients $\rho = corr(u, v)$.

To solve this problem, we create a matrix P whose rows are the vectors $\Phi_x(X_i), i = 1, \ldots, m$, and similarly a matrix Q with rows $\Phi_y(Y_i), i = 1, \ldots, m$.

The covariance matrix of P and Q can be described as:[3]

$$C_{pp} = P^T P. \tag{9}$$

$$C_{pq} = P^T Q. \tag{10}$$

[2] For more details, please refer to Statistical Consistency of Kernel Canonical Correlation (Fukumizu et al. 2007)

[3] P and Q are normalized matrix.

Further more, the projection vector a and b in can be expressed as a linear combination of the training examples:[4]

$$a = P^T \alpha. \tag{11}$$

$$b = Q^T \beta. \tag{12}$$

where α, β are m-dimensional vectors as the parameter of kernel CCA. So far, the object to maximize can be described as:

$$\rho(u,v) = \frac{cov(u,v)}{\sqrt{Var(u)}\sqrt{Var(v)}} = \frac{\alpha^T PP^T QQ^T \beta}{\sqrt{\alpha^T PP^T PP^T \alpha}\sqrt{\beta^T QQ^T QQ^T \beta}} \tag{13}$$

Given the kernel functions Φ_x and Φ_y, let K_p and K_q be the kernel matrices corresponding to two representations of the data, where $K_p = PP^T, K_q = QQ^T$. Substituting it into Equation (13), we get:

$$\rho(u,v) = \frac{\alpha^T K_p K_q \beta}{\sqrt{\alpha^T {K_p}^2 \alpha}\sqrt{\beta^T {K_q}^2 \beta}} \tag{14}$$

Observing that Equation (14) is not affected by the rescaling of α and β either together or independently, the kernel CCA optimization problem is equivalent to maximizing the numerator subject to:

$$\alpha^T {K_p}^2 \alpha = \beta^T {K_q}^2 \beta = 1 \tag{15}$$

The corresponding Lagrangian with regularization for KCCA is:

$$L = \alpha^T K_p K_q \beta - \frac{\lambda}{2}(\alpha^T K_p^2 \alpha - 1) - \frac{\theta}{2}(\beta^T K_q^2 \beta - 1) + \frac{\eta}{2}(\| \alpha \|^2 + \| \beta \|^2) \tag{16}$$

By solving Equation (16), α and β can be found by an eigenvector corresponding to the maximal eigenvalues of a generalized eigenvalue problem:[5]

$$(K_p + \eta I)^{-1} K_q (K_q + \eta I)^{-1} K_p \alpha = \lambda^2 \alpha \tag{17}$$

Thus we have found the first pair of canonical variables. If more dimension is needed, we can take eigenvectors corresponding to other maximal eigenvalues.

3.2 Generating Multi-lingual Embeddings Using KCCA

Now we describe how to apply KCCA to our task. We first use KCCA to learn the relationship between two monolingual vocabularies from a bilingual lexicon and output two transformation matrices. By using these two transformation matrices, we then project the original word embeddings onto a new shared vector space. The schema of performing KCCA on the monolingual word representations of two languages (BiKCCA) is shown in Figure. 3.

Let $\Sigma \in \mathbb{R}^{n_1 \times d_1}$ and $\Omega \in \mathbb{R}^{n_2 \times d_2}$ be word embeddings of two languages vocabularies, where n_1, n_2 represent the size of vocabularies, d_1, d_2 denote the dimensionality of word vectors. The training matrices Σ', Ω' are constructed as $\Sigma' \subset \Sigma$, $\Omega' \subset \Omega$, where Ω'_i is translated from Σ'_i.

We use KCCA to maximize ρ for the given set Σ' and Ω' and output two parameters α, β:

$$\alpha, \beta = KCCA(\Sigma', \Omega') = \arg\max_{\alpha,\beta} \rho(a^T \Phi_x(\Sigma'), b^T \Phi_y(\Omega')) \tag{18}$$

a, b in Equation (18) can be denoted by α, β as Equation (11), (12).

[4] This has been proved by (Fukumizu et al. 2007), by expressing a and b like that, kernel trick could be used.

[5] η is a known variable , I is identity matrix.

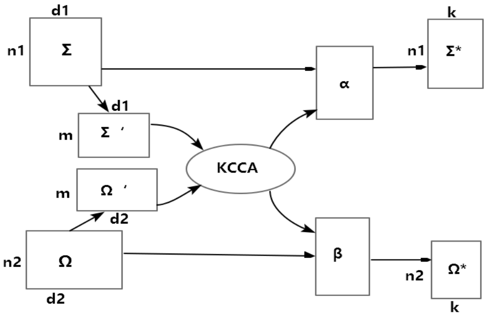

Fig. 3. Generating cross-lingual word vector using KCCA

Using such two vectors α, β, we can project the entire vocabulary of the two languages Σ and Ω onto a shared vector space $\mathbb{R}^{D_k}$ via Equation (7), (8). Substituting into Equation (11), (12), it could be summarized as:

$$\alpha_k, \beta_k = KCCA(\Sigma', \Omega') \tag{19}$$

$$\Sigma^*_k = P_\Sigma P^T_{\Sigma'} \alpha_k, \Omega^*_k = Q_\Omega Q^T_{\Omega'} \beta_k \tag{20}$$

where P_Σ is a matrix whose rows are vectors $\Phi_x(X_i)$[6], $X_i \in \Sigma$, Q_Ω is a matrix whose rows are vectors $\Phi_y(Y_i), Y_i \in \Omega$, and $P_{\Sigma'}, Q_{\Omega'}$ are similar to P_Σ, Q_Ω. We perform experiments by taking projections of the top k correlated dimensions, α_k, β_k are the corresponding parameter matrix. $\Sigma^*_k \in \mathbb{R}^{n_1 \times k}$, $\Omega^*_k \in \mathbb{R}^{n_2 \times k}$ are new word vector matrices for two languages.

4 WORD REPRESENTATION EVALUATION

Upadhyay et al. (2016) performed a systematic comparison of four popular approaches of inducing cross-lingual embeddings. We follow the setup of Upadhyay to evaluate our embeddings. The following three tasks are performed to assess the quality of the induced cross-lingual word embeddings:

- Monolingual word similarity for English
- Cross-lingual dictionary induction
- Cross-lingual document classification

The first two tasks intrinsically show how much benefit would be gained from cross-lingual training. The last task measures the ability of cross-lingually trained vectors to extrinsically facilitate model transfer across languages.[7]

4.1 Monolingual Evaluation

We first evaluate whether the proposed model can improve the quality of English embeddings. The task of monolingual evaluation assesses how well the trained word embeddings capture human intuitions about semantic relatedness via word similarity datasets.

We use the SimLex dataset for English (Hill et al. 2014) which contains 666 noun pairs, 222 verb pairs, and 111 adjective pairs. SimLex is claimed to capture word similarity merely instead of WordSim-353 (Finkelstein et al. 2001) which captures both word similarity and relatedness. To evaluate embeddings, we compute cosine similarity between two vectors in each pair, order the

[6] Φ_x, Φ_y are kernel function defined in §3.1

[7] To ensure a fair comparison, both models are trained with embeddings of size200, and product embeddings of size100.

pairs by similarity, and compute Spearman correlation (ρ) (Myers and Well 1995) between the model's ranking and human ranking.

4.2 Cross-lingual Dictionary Induction

In this task, we assess whether our non-linear model improves the quality of the vector on cross-lingual dictionary induction task. The task of cross-lingual dictionary induction (Vulić and Moens 2013; Gouws et al. 2015; Mikolov et al. 2013b; Upadhyay et al. 2016; Zhang et al. 2016) judges the ability of cross-lingual embeddings to detect word pairs that are semantically similar across languages.

We follow the setup of Upadhyay, and derive our gold dictionaries via the Open Multilingual WordNet data released by Bond and Foster (2013). Using the approach described in Upadhyay's paper, we generated dictionaries of sizes 1.5k,1.4k,1.5k,1.4k and 1.6k pairs for en-fr, en-de, en-ar, en-ru and en-zh respectively. Top-k ($k \in \{1, 10, 50\}$) accuracy is reported on this task. For each pair (e, f) in the gold dictionary, we check if f belongs to the list of top-k neighbors of e, according to the induced cross-lingual word vectors.

4.3 Cross-lingual Document Classification

The cross-lingual document classification task assesses whether the learned cross-lingual representations are semantically coherent across multiple languages.

We follow the cross-lingual document classification (CLDC) setup of Upadhyay et al. (2016), but choose another popular corpus WIT TED (Cettolo et al. 2012) which has more topics available than RCV2 (Lewis et al. 2004) for our evaluation. 15 topics are chosen for the classification task in our experiment. Document representations are computed by taking the tf-idf weighted average of vectors of the words present in it.[8] A multi-class classifier is then trained on the labeled training data in the source language via an averaged perceptron (Freund and Schapire 1999) for 10 iterations, using the document vectors of language l_1 as features. For each language pair (l_1, l_2), we train a document classifier using the document embeddings derived from word embeddings in language l_1, and test this model on document embeddings from l_2.

We reference the result of MT system trained by Hermann and Blunsom on TED corpus as a baseline. It uses the cdec decoder (Dyer et al. 2010) with default settings to translate documents.

5 EXPERIMENTS

In this section we perform the tasks described in (§4) to evaluate the utility of the induced bilingual word embeddings and present the results.

5.1 Data

We train cross-lingual embeddings for 5 language pairs: English-German (en-de), English-French (en-fr), English-Arabic (en-ar), English-Russian (en-ru) and English-Chinese (en-zh). For English, French, German, monolingual corpora are obtained from Europarl,[9] and for Arabic, Russian, Chinese, text from Leipzig[10] is used.

To generate a bilingual lexicon, parallel corpora from Europarl are used. First, word pair (a, b), $a \in l_1$, $b \in l_2$ is selected such that a is aligned to b the most number of times in parallel corpus and vice versa. Then, our bilingual dictionary is constructed from the most common pairs. For Arabic, Russian, Chinese where parallel corpus is unavailable in Europarl, dictionaries are induced by

[8] tf-idf (Salton and Buckley 1988) was computed by using all documents for each language in TED.

[9] http://www.statmt.org/europarl/

[10] http://wortschatz.uni-leipzig.de/en

Table 1. Intrinsic evaluation of English word vectors

L_1	L_2	Mono	BiCCA	BiKCCA
en	fr	0.291	0.303	**0.305**
	de	0.297	0.312	**0.318**
	ar	0.283	0.295	**0.327**
	ru	0.281	0.301	**0.317**
	zh	0.283	0.303	**0.322**
avg.		0.287	0.303	**0.318**

Word similarity score measured in Spearman's correlation ratio for English on SimLex-999, with higher being better. Scores which are significantly better (Steiger's Method with p < 0.15) are underlined. Bold indicates best result in each language pair. The similar trend was also observed when computing QVEC (Tsvetkov et al. 2015).

translating the 20k most common words in the English monolingual corpus with Google Translation. To balance the speed and performance, we use a bilingual lexicon about 7k words for each language pair to train model in practice.

5.2 Settings

First, original monolingual vectors are trained via the skip-gram model[11] with negative sampling (Mikolov et al. 2013a) with window of size 5 (tuned over 5, 10, 20). Bilingual dictionaries are generated as described in (§5.1). We use k = 0.5 as the scaling factor of canonical components (tuned over 0.2, 0.3, 0.5, 1.0) as also done by Faruqui and Dyer. After performing this, we get embeddings of size 100 for evaluation.

For KCCA, we use a radial basis function (RBF) kernel[12] for both views: $k_1(x_i, y_i) = exp(-\gamma(\| x_i - y_i \|^2)$ and similarly for k_2. The parameters γ_1, γ_2 are tuned over the range $[10^{-1}, 10^{-5}]$. Regularization parameter η in Equation (15) is tuned over the range $[10, 10^{-6}]$. For each evaluation task, we perform 5-fold cross-validation and choose the hyperparameters with the best performance.

5.3 Results

Monolingual Word Similarity Table 1 shows our main results of the monolingual word similarity task (§4.1). We compare the performance of monolingual word embeddings, CCA-based bilingual word embeddings(BiCCA) and KCCA-based bilingual word embeddings(BiKCCA). Monolingual word embedding is used as a baseline. We declare significant improvement if p < 0.15 according to Steiger's method (Steiger 1980) for calculating the statistical significant differences between two dependent correlation coefficients.

It can be seen from Table 1 that BiKCCA results show consistent improvements across most language pairs over BiCCA (all the BiKCCA results in bold are better than BiCCA). This indicates that the non-linear relationship is useful to improve the quality of word embeddings on word similarity task. Moreover, we note that across language pairs which are substantially different, such as English-Russian, English-Arabic, English-Chinese, BiKCCA achieves greater improvement over BiCCA than other language pairs.

Cross-lingual Dictionary Induction In Table 2, we report the results of cross-lingual dictionary induction task (§4.2). Top-k ($k \in \{1, 10, 50\}$) accuracy is presented, for each pair (e, f) in the gold dictionary, we check if f belongs to the list of top-k neighbors of e, according to the induced cross-lingual word vectors.

[11] code.google.com/p/word2vec

[12] We also tried poly kernel, did not observe any superiority in performance.

Table 2. Cross-lingual dictionary induction

L_1	L_2	BiCCA			BiKCCA		
		top-1	top-10	top-50	top-1	top-10	top-50
en	fr	52.4	70.3	79.8	**53.3**	**71.1**	**79.9**
	de	53.2	72.4	80.4	**54.6**	**72.9**	**80.6**
	ar	32.7	53.1	62.1	**49.8**	**66.1**	**72.8**
	ru	37.3	55.4	63.4	**49.9**	**66.6**	**74.4**
	zh	36.3	60.1	73.1	**46.1**	**66.7**	**76.8**
avg.		42.4	62.3	71.8	**50.8**	**68.7**	**76.9**

Cross-lingual dictionary induction results (top-k accuracy,k={1,10,50}). Bold indicates best result. The similar trend was also observed across models when computing MRR (mean reciprocal rank)

Table 3. Cross-lingual Document Classification results

Setting	Languages				
	French	German	Arabic	Russian	Chinese
En-L_2					
MT Baseline	**0.526**	**0.465**	**0.429**	**0.432**	-
BiCCA	0.446	0.399	0.344	0.276	0.204
BiKCCA	0.446	0.410	0.345	0.285	0.223
L_2-En					
MT Baseline	0.358	**0.469**	**0.448**	**0.404**	-
BiCCA	0.452	0.387	0.373	0.329	0.374
BiKCCA	**0.472**	0.435	0.387	0.346	0.375

F1-scores for the TED document classification task for individual languages. Results are for two directions (training on English, evaluating on L2 and vice versa). Bold indicates best result, underline indicates best result between the bilingual vectors. We refer the MT Baseline reported by Hermann and Blunsom.

From Table 2, we note that by capturing non-linear relationship between two source languages, BiKCCA performs essentially better than BiCCA across all language pairs (all of the bold BiKCCA results are better than all corresponding CCA results). This matches our assumption that non-linear model is more suitable for capturing factors among languages. Clearly, the effect of BiKCCA on substantially different language pairs like en-ar, en-ru, en-zh, is much more obvious.

Cross-lingual Document Classification Table 3 shows performance of different models on cross-lingual document classification task (§4.3) across different language pairs. We report average F1-score of 15 topics, which can be interpreted as a weighted average of the precision and recall. The 4th and 8th row of Table 3 show the result of MT system reported by Hermann and Blunsom. We list it here for reference but note that it is not comparable to our results since this system has access to significantly more information as opposed to our models, and we do not expect to defeat this system.

When comparing the results of the linear model(BiCCA) and non-linear model(BiKCCA), the benefit of this additional non-linear relationship becomes clear (all the BiKCCA results underlined are better than BiCCA). This suggests that for transferring semantic knowledge across languages via embeddings, non-linear model proves superior to linear model.

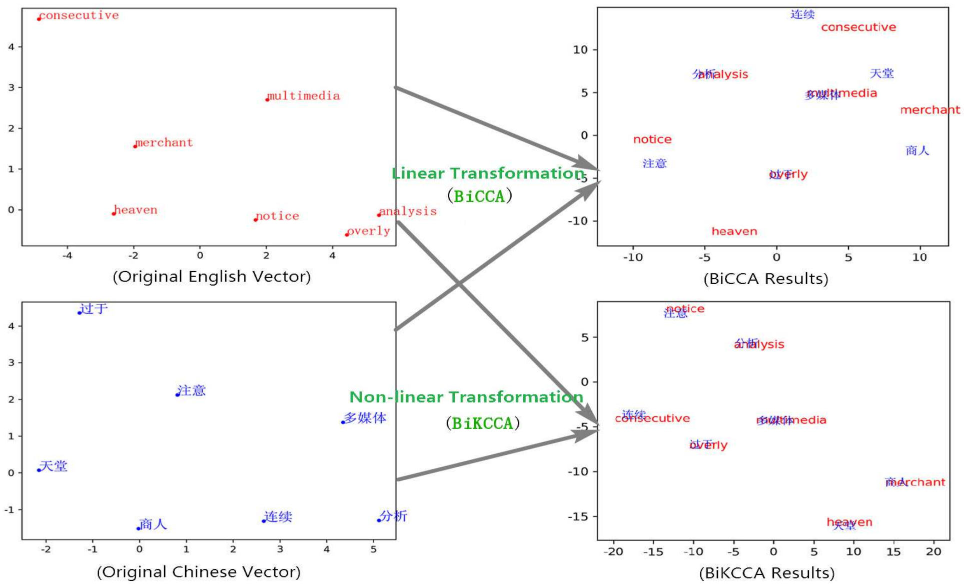

Fig. 4. t-SNE visualization of some frequent words in English-Chinese corpus. English and Chinese words are shown in red and blue respectively.

Table 4. Cosine similarity between the two vectors in each word pair

word pair	Linear Trans (BiCCA)	Non-linear Trans (BiKCCA)
(analysis, 分析)	0.706	0.883
(multimedia, 多媒体)	0.863	0.935
(consecutive, 连续)	0.685	0.821
(heaven, 天堂)	0.656	0.830
(merchant, 商人)	0.622	0.821

6 ANALYSIS

Result Analysis To further understand how BiKCCA gets better performance in our evaluation tasks, we analyze the distribution of the cross-lingual word embeddings in shared vector space. Figure 4 shows the t-SNE (Maaten and Hinton 2008) visualization of some high frequency word pairs in the English-Chinese corpus. The original monolingual word vectors as well as bilingual word vectors generated by BiCCA (linear transformation) and BiKCCA (non-linear transformation) are presented in left and right side respectively. For each word pair, our goal is to learn transformations for each view to make two words across languages have similar representations, which can also be described as two words are right near in vector space. By comparing these two regions, we observe that (i) For word pairs as (heaven, 天堂), (merchant, 商人), BiKCCA catches the corresponding words in Chinese while BiCCA doesn't (The distance between these words is too long in BiCCA). (ii) For word pairs like (notice, 注意), (consecutive, 连续) which are caught both by BiCCA and BiKCCA, the distance between two translationally equivalent words in BiKCCA's vector space

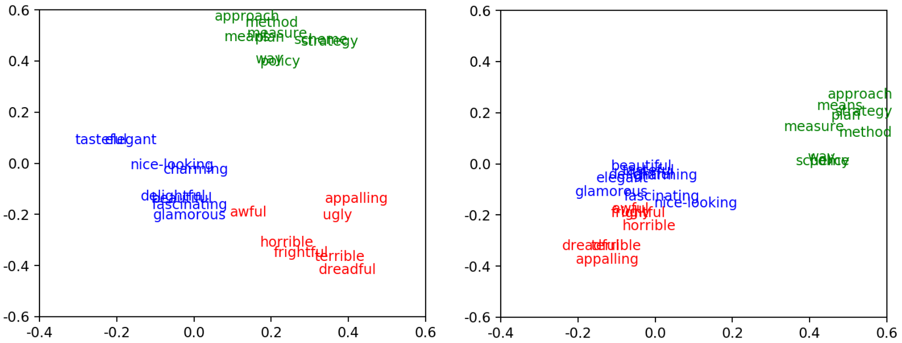

Fig. 5. PCA projection of word embeddings of three groups of near-synonym in English,different groups with different colors. BiCCA in left and BiKCCA in right.

is shorter than BiCCA, which is a good property in many downstream applications. These two observations explain how BiKCCA gets higher accuracy on cross-lingual dictionary induction and cross-lingual document classification tasks.

We also present the cosine similarity between vectors[13] of each word pair to verify our observations in Table 4, higher is better. By comparing the result of two models, we found that BiKCCA leads to a relative improvement over BiCCA (cosine similarity between vectors in BiKCCA are higher than BiCCA).

The reason for these two observations described above is that the relationship between words of different languages can't be captured well by a linear model, and this results in longer distance between two translationally equivalent words, while our non-linear model is able to group translationally equivalent words together in the vector space.

In Figure 5, we present the PCA projection of three groups of synonym in English (embeddings are trained on En-Zh corpus). It can be easily seen that synonyms in red and blue group of right region (BiKCCA) are closer than left region (BiCCA), and both models perform well in green group. This indicates that semantically similar words in the same language in BiKCCA are closer than BiCCA in most groups, and this also explains the better performance of BiKCCA on word similarity task. The reason for this phenomenon is that for English-side transformation, non-linear transformation reflects the fact better than linear transformation. In addition, we have also found that both models are good at separating antonyms (The red group and blue group is well separated in BiCCA and BiKCCA). This can be attributed to the fact that both models use bilingual dictionary, which helps in pulling apart the synonyms and antonyms.

Last but not least, basing on results of three evaluation tasks, we discover that in closely related languages, BiKCCA gains little enhancement in performance over BiCCA whether on intrinsic or extrinsic tasks, while in substantially different languages, BiKCCA gains great improvement. Supposing that we have captured almost all non-linear relationships,[14] from a quantitative point of view, we could draw a conclusion that there are a lot of non-linear relationships among substantially different languages, while little among closely related languages.

[13] vectors have been normalized before computation

[14] Actually the maximum correlation coefficient ρ has reached 0.99 in our experiment.

Error Analysis We analyze word pairs that change the most with BiCCA and BiKCCA on cross-lingual dictionary induction task. A typical error is that a word is often mistranslated into another word with close but different meaning, for example word “way/方法” is mistranslated into “政策/policy”. This error appears in both models, with a lot in BiCCA and less in BiKCCA. According to Figure 5, we observe that semantically similar words in the same language always gather together, words like "way", "policy" are gathered into one group, the same to “方法” , “政策” . However, we have also shown that translationally equivalent words are also close in shared vector space, this causes that all semantically similar words crowd together so that word is often mistranslated. BiKCCA avoids part of such error because translationally equivalent words are so close in vector space that noise words make little sense, while BiCCA can’t reduce this error due to the limitation of linear transformation.

Model Analysis We have proposed a model which is computationally-efficient, easy to scale to large datasets and outperforms CCA-based model in chosen evaluation tasks. But our model also has two limitations — one is that it is not able to learn multiple embeddings per word. Actually, homonymy and polysemy are common in natural languages, and they are often the sources of error in word embedding algorithms. The other shortcoming of our method is that it only learns embeddings at word level, hence currently we can not capture compositional semantics.

7 CONCLUSIONS AND FUTURE WORK

We have presented a KCCA-based approach for learning multilingual word embeddings, which addresses drawbacks of CCA-based model. Basing on results from extrinsic and intrinsic evaluation tasks, we show that KCCA embeddings consistently outperform CCA embeddings on each task across various language pairs, especially across substantially different languages. By qualitatively analyzing, we note that previous linear assumption does not always hold true, and non-linear assumption is more suitable for representing the relationship among languages. In the future, we would extend our experiment to more languages to further inspect the performance of our model, and apply our method to more downstream applications such as cross-lingual dependency parsing and machine translation.

REFERENCES

Shotaro Akaho. 2006. A kernel method for canonical correlation analysis. *CoRR* abs/cs/0609071 (2006).

Francis Bond and Ryan Foster. 2013. Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 1352–1362.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)* (28-30). Trento, Italy, 261–268.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-Based Dependency Parsing with Stack Long Short-Term Memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, 334–343.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Jonathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A Decoder, Alignment, and Learning Framework for Finite-State and Context-Free Translation Models. In *Proceedings of the ACL 2010 System Demonstrations*. Association for Computational Linguistics, Uppsala, Sweden, 7–12.

Manaal Faruqui and Chris Dyer. 2014. Improving Vector Space Word Representations Using Multilingual Correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Gothenburg, Sweden, 462–471.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing Search in Context: The Concept Revisited. In *Proceedings of the 10th International Conference on World Wide Web (WWW ’01)*. ACM, New York, NY, USA, 406–414.

Yoav Freund and Robert E. Schapire. 1999. Large Margin Classification Using the Perceptron Algorithm. *Machine Learning* 37, 3 (1999), 277–296.

Kenji Fukumizu, Francis R. Bach, and Arthur Gretton. 2007. Statistical Consistency of Kernel Canonical Correlation Analysis. *J. Mach. Learn. Res.* 8 (May 2007), 361–383.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, David Blei and Francis Bach (Eds.). JMLR Workshop and Conference Proceedings, 748–756.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual Dependency Parsing Based on Distributed Representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, 1234–1244.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Models for Compositional Distributed Semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 58–68.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *CoRR* abs/1408.3456 (2014).

Harold Hotelling. 1936. Relations Between Two Sets of Variates. *Biometrika* 28, 3/4 (1936), 321–377.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing Crosslingual Distributed Representations of Words. In *COLING*.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A New Benchmark Collection for Text Categorization Research. *J. Mach. Learn. Res.* 5 (Dec. 2004), 361–397.

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*. 151–159.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 92–97.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013).

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting Similarities among Languages for Machine Translation. *Computer Science* (2013).

Jerome L Myers and A. (Arnold) Well. 1995. *Research design and statistical analysis* (1nd ed ed.). Mahwah, N.J. : Lawrence Erlbaum Associates. Includes bibliographical references (p. 729-741) and indexes.

Gerard Salton and Christopher Buckley. 1988. Term-weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manage.* 24, 5 (Aug. 1988), 513–523.

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)*. Association for Computational Linguistics, Beijing, China, 567–572.

Richard Socher, John Bauer, Christopher D. Manning, and Ng Andrew Y. 2013. Parsing with Compositional Vector Grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 455–465.

James H Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin* 87, 2 (1980), 245.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of Word Vector Representations by Subspace Alignment. In *EMNLP*.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual Models of Word Embeddings: An Empirical Comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1661–1670.

Jakob Uszkoreit, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner. 2010. Large Scale Parallel Document Mining for Machine Translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1101–1109.

Ivan Vulić and Marie-Francine Moens. 2013. Cross-Lingual Semantic Similarity of Words as the Similarity of Their Semantic Word Responses. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, 106–116.

Xiaojun Wan. 2009. Co-training for Cross-lingual Sentiment Classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1 (ACL '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 235–243.

Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, and Chengqing Zong. 2014. Bilingually-constrained Phrase Embeddings for Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 111–121.

Meng Zhang, Yang Liu, Huanbo Luan, Maosong Sun, Tatsuya Izuha, and Jie Hao. 2016. Building Earth Mover's Distance on Bilingual Word Embeddings for Machine Translation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16)*. AAAI Press, 2870–2876.