Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3583–3589
Marseille, 11–16 May 2020

Non-Linearity in mapping based Cross-Lingual Word Embeddings

Jiawei Zhao, Andrew Gilman
Massey University
Auckland, New Zealand
{j.zhao1, a.gilman}@massey.ac.nz

Abstract

Recent works on cross-lingual word embeddings have been mainly focused on linear-mapping-based approaches, where pre-trained word embeddings are mapped into a shared vector space using a linear transformation. However, there is a limitation in such approaches–they follow a key assumption: words with similar meanings share similar geometric arrangements between their monolingual word embeddings, which suggest that there is a linear relationship between languages. However, such assumption may not hold for all language pairs across all semantic concepts. We investigate whether non-linear mappings can better describe the relationship between different languages by utilising kernel Canonical Correlation Analysis (KCCA). Experimental results on five language pairs show an improvement over current state-of-art results in both supervised and self-learning scenarios, confirming that non-linear mapping is a better way to describe the relationship between languages.

Keywords: Cross-lingual Word Embedding, KCCA, Non-linearity

1. Introduction

Cross-lingual representations have gained much interest recently. It has been shown that cross-lingual word embedding models succeed in many inherently cross-lingual Natural Language Processing (NLP) tasks such as machine translation and cross-lingual entity linking (Artetxe et al., 2018c)(Lample et al., 2018)(Tsai and Roth, 2016). Cross-lingual word embedding models also allow to reason word semantics in multi-context environments and helps to transfer semantic knowledge from rich- to low-resource languages.

Linear-mapping-based cross-lingual word embeddings rely on a basic assumption, first stated by Mikolov et al. (2013), that words with similar meanings have a similar geometric arrangement in the embedding vector space. However, this assumption may be heavily violated for languages with different cultural backgrounds. Therefore, we hypothesis that non-linearity can better express the relationships between languages. Based on such hypothesis, we perform a non-linear alignment between word embeddings of different languages.

In this paper, we introduce a KCCA-based mapping approach, which can find non-linear relationships between languages. Our experiments show that the addition of non-linearity can improve cross-lingual word embeddings for a number of language pairs. Our code and the new English-Chinese dataset are released under open source license[1].

We review prior work in Section 2. In Section 3, we introduce our proposed approach. Section 4 presents our experiments and analysis of the experimental results is given in Section 5. Section 6 concludes our work.

2. Related Work

The mapping-based approach was first proposed by Mikolov et al. (2013), who first train monolingual word embeddings for two languages independently, and then map those embeddings into a shared vector space based by minimising the Euclidean distance between embeddings representing the same word in the two languages. Xing et al. (2015) also followed this approach, but argued that word embeddings should be normalised to unit length. Artetxe et al. (2016) further showed that the projection matrices should be constrained to orthogonal. Artetxe et al. (2018b) proposed a general framework to unify existing linear-mapping-based approaches and performed an empirical comparison between existing methods.

An alternative approach is to maximise the correlation between words in different languages, which can be learnt using Canonical Correlation Analysis (CCA). Faruqui and Dyer (2014) were the first to apply CCA to construct cross-lingual word representations from two sets of monolingual ones and demonstrated that their use (instead of monolingual representations) improved the performance of several tasks. They attribute the improved performance to the idea that shared representation is able to incorporate lexico-semantic information from both languages. Ammar et al. (2016) extended this work to a multi-lingual scenario, which is able to project word vectors from more than two languages into a shared embedding space.

Artetxe et al. (2016) showed that Euclidean distance-based approach and correlation-based approach are inherently the same, except that CCA imposes a constraint of equal variance in each component of the new word representations. They argue that this kind of constraint may have a negative impact on the performance, but their evaluation on the word translation task, as well as, our own experiments do not support this claim.

The only proposed approach, capable of mapping non-linear relationships is that of Lu et al. (2015), based on Deep Canonical Correlation Analysis (DCCA). DCCA have mainly two weaknesses. Firstly, DCCA utilises deep neural networks with many parameters, which requires a large amount of high-quality training data. Secondly, the method also requires tuning a large number of hyperparameters, which in our experience can be dataset-specific

[1] https://gitlab.com/zjw1990/kclwe

and difficult to optimise. In contrast, our proposed KCCA-based model only has a few hyperparameters to tune, which makes our model simpler to use in practice. Also, it does not require large amounts of training data. This is confirmed by the result our experiments: KCCA-based model works better than DCCA.

3. Method

3.1. CCA and KCCA

First, we briefly introduce CCA. Given two multivariate random variables (i.e. two random column vectors) $a \in \mathbb{R}^{d_a}$ and $b \in \mathbb{R}^{d_b}$, CCA aims to find basis vectors $w_a \in \mathbb{R}^{d_a}$ and $w_b \in \mathbb{R}^{d_b}$ such that the correlation, ρ, between projections onto these basis $w_a^\intercal a$ and $w_b^\intercal b$ is mutually maximised:

$$w_a, w_b = \underset{w_a, w_b}{\arg\max}\, \rho(w_a^\intercal a, w_b^\intercal b) \tag{1}$$

where $\bullet^\intercal$ denotes the transpose operator. In the case where each dimension of a and b is centred, i.e. $\mathbb{E}[a_0] = ... = \mathbb{E}[a_{d_a-1}] = \mathbb{E}[b_0] = ... = \mathbb{E}[b_{d_b-1}] = 0$, this optimisation can be expressed as:

$$w_a, w_b = \underset{w_a, w_b}{\arg\max} \frac{w_a^\intercal C_{ab} w_b}{\sqrt{w_a^\intercal C_{aa} w_a}\sqrt{w_b^\intercal C_{bb} w_b}} \tag{2}$$

where C_{ab}, C_{aa} and C_{bb} denote covariance matrices. Since scaling w_a and w_b has no effect on equation 1, this is equivalent to maximising the numerator subject to an additional constraint $w_a^\intercal C_{aa} w_a = w_b^\intercal C_{bb} w_b = 1$. Such optimisation can be formulated as the Lagrangian and solved through Lagrangian relaxation.

The main limitation of CCA is its linearity. In contrast, KCCA first projects the data into a higher-dimensional space using a mapping function ϕ:

$$\phi : \boldsymbol{p} = (p_1, \ldots p_d) \mapsto \phi(\boldsymbol{p}) = (\phi_1(\boldsymbol{p}), \ldots, \phi_D(\boldsymbol{p}))\ (d \ll D) \tag{3}$$

that maps random variable $\boldsymbol{p}$ from the original space $\mathbb{R}^d$ to a new space $\mathbb{R}^D$; and then performs CCA in this new feature space (Lai and Fyfe, 2000). Kernel methods are algorithms, widely used in machine learning, that do not require one to specify ϕ explicitly; instead, only a kernel function that allows computing the inner product of two data points in the new feature space needs to be specified:

$$k(\boldsymbol{p}, \boldsymbol{p'}) = \langle \phi(\boldsymbol{p}), \phi(\boldsymbol{p'}) \rangle \tag{4}$$

Now, any machine learning algorithm that can be expressed via inner products, can be computed in the new high-dimensional feature space, without explicitly projecting the data or even knowing the mapping function ϕ.

In practice, we only have a sample of instances of the random vectors a and b. Consider data matrices $A \in \mathbb{R}^{N \times D}$ and $B \in \mathbb{R}^{N \times D}$, whose rows contain the sample vectors in the new high-dimensional feature space. We can rewrite equation 2 by expressing the covariance matrices in terms of these data matrices ($C_{aa} = A^\intercal A$, $C_{bb} = B^\intercal B$, $C_{ab} = A^\intercal B$):

$$\underset{w_a, w_b}{\arg\max} \frac{w_a^\intercal A^\intercal B w_b}{\sqrt{w_a^\intercal A^\intercal A w_a}\sqrt{w_b^\intercal B^\intercal B w_b}} \tag{5}$$

We can express the basis w_a and w_b as linear combinations of the data points using coefficients $\alpha \in \mathbb{R}^N$ and $\beta \in \mathbb{R}^N$:

$$w_a = A^\intercal \alpha \tag{6}$$

$$w_b = B^\intercal \beta \tag{7}$$

Then the dual representation of the problem can be formulated by substituting Eq. 6,7 into Eq. 5:

$$\underset{\alpha, \beta}{\arg\max} \frac{\alpha^\intercal A A^\intercal B B^\intercal \beta}{\sqrt{\alpha^\intercal A A^\intercal A A^\intercal \alpha}\sqrt{\beta^\intercal B B^\intercal B B^\intercal \beta}} \tag{8}$$

Let $K_a = AA^\intercal$ and $K_b = BB^\intercal$ be the kernel matrices (Gram matrices), substituting these into Eq. 8, we get:

$$\underset{\alpha, \beta}{\arg\max} \frac{\alpha^\intercal K_a K_b \beta}{\sqrt{\alpha^\intercal K_a^2 \alpha}\sqrt{\beta^\intercal K_b^2 \beta}} \tag{9}$$

Hardoon et al. (2004) observed that KCCA frequently suffers over-fitting, especially when dealing with high-dimensional data and applied regularisation to reduce this:

$$\underset{\alpha, \beta}{\arg\max} \frac{\alpha^\intercal K_a K_b \beta}{\sqrt{\alpha^\intercal K_a^2 \alpha + \kappa \left\| w_a \right\|^2}\sqrt{\beta^\intercal K_b^2 \beta + \kappa \left\| w_b \right\|^2}} \tag{10}$$

It follows that:

$$\underset{\alpha, \beta}{\arg\max} \frac{\alpha^\intercal K_a K_b \beta}{\sqrt{\alpha^\intercal K_a^2 \alpha + \kappa \alpha^\intercal K_a \alpha}\sqrt{\beta^\intercal K_b^2 \beta + \kappa \beta^\intercal K_b \beta}} \tag{11}$$

where κ controls the amount of regularisation that is applied. Similarly to CCA, since this problem is not affected by scaling of α and β, it can be reformulated as a maximisation of the numerator, subject to

$$\left(\alpha^\intercal K_a^2 \alpha + \kappa \alpha^\intercal K_a \alpha\right) = 1 \tag{12}$$

$$\left(\beta^\intercal K_b^2 \beta + \kappa \beta^\intercal K_b \beta\right) = 1 \tag{13}$$

Through the Lagrangian formulation, this leads to a standard eigenproblem:

$$\left(K_a + \kappa I\right)^{-1} K_b \left(K_b + \kappa I\right)^{-1} K_a \alpha = \lambda^2 \alpha \tag{14}$$

The eigenvalues of Eq.14 are the canonical correlations and the eigenvectors can be used to calculate the projections. This problem can be solved in different ways; however, we choose an effective algorithm (PGSO) proposed by Hardoon et al. (2004). We reproduce their Matlab implementation[2] in Python.

[2] https://davidroihardoon.com/codes

3.2. KCCA based cross-lingual word embedding model

Let $X \in \mathbb{R}^{N_x \times d_x}$ and $Y \in \mathbb{R}^{N_y \times d_y}$ be monolingual word embeddings from vocabularies of the source and target languages. In such embedding matrices, rows represent words and columns represent features of words. In the supervised scenario, a set of word embeddings of translation pairs (i.e a dictionary) is given: let $x \in \mathbb{R}^{n \times d_x}$ contain a subset of embeddings from X and $y \in \mathbb{R}^{n \times d_y}$ contain their translations from Y, such that the same row in each matrix represents a translation pair. Our proposed approach can be summarised into three steps: pre-processing, KCCA-projection and re-weighting based on canonical correlations.

3.2.1. Step 1: Pre-processing:

Before applying KCCA, source and target word embeddings are pre-processed with length normalisation and mean centering. Length normalisation is applied sample-wise, such that all embeddings have Euclidean unit length. Mean centering makes all components have a zero mean.

3.2.2. Step 2: KCCA Projection:

This step contains 2 parts. First, learn projections using the dictionary, then use those learnt projections to project the vocabulary into the new space. Specifically:

Learn Mapping Given word embedding matrices x and y, as defined above, we adopt the KCCA implementation described in Section 3.1.:

$$\alpha, \beta, \lambda = KCCA(x, y) \tag{15}$$

where α and β are the coefficient vectors, described in Eq. 6, 7), and λ is the canonical correlations corresponding to each of the projection directions. We use the Radial Basis Function (RBF) kernel and tune the value of parameter γ through cross-validation.

Vocabulary Projection Given α and β calculated by KCCA, the vocabularies X and Y are projected into the shared space:

$$X^* = \phi(X) \cdot w_x \tag{16}$$

Substituting w_x with Eq. 6:

$$X^* = \phi(X) \cdot \phi(x^\intercal) \cdot \alpha \tag{17}$$

Then the inner product $\phi(X) \cdot \phi(x^\intercal)$ can be expressed using the kernel trick:

$$K(X, x^\intercal) = \langle \phi(X) \cdot \phi(x^\intercal) \rangle \tag{18}$$

Substitute Eq.17 with Eq.18, the new representation of the vocabulary X^* then can be calculated as:

$$X^* = K(X, x^\intercal) \cdot \alpha \tag{19}$$

And similarly for Y^*.

3.2.3. Step 3: Canonical Correlation Re-weighting:

The re-weighting process is described by Artetxe et al. (2018a). After projection, the components of the new embeddings are re-weighted based on their singular values, which can be used to increase the strength of relations that have best matched across languages. However, they failed to make re-weighting work with CCA and did not use it with CCA. We were able to successfully adapt the process and apply it to KCCA. The components of the new embeddings are re-weighted based on their canonical correlations:

$$X^* = X^* \lambda^\zeta \tag{20}$$

$$Y^* = Y^* \lambda^\zeta \tag{21}$$

where ζ is a tune-able parameter with a default value of 1; however, different language pairs may require slightly different value of this parameter to get optimal results and ideally it should be tuned.

4. Experiments

In our experiments, we aim to investigate whether non-linear mapping is better than linear mapping for producing cross-lingual word embeddings. We evaluate CCA-, DCCA- and KCCA-mapped word embeddings on the word translation task with different languages. Also, we compare our result with other linear-mapping-based approaches.

4.1. Dataset and Task

We use four language pairs in our experiments: English-Italian, English-German, English-Spanish and English-Finnish. The English-Italian dataset is provided by Dinu et al. (2015) and extended to other three language pairs by Artetxe et al. (2017). Each dataset includes 20k 300 dimensional monolingual word embeddings trained by word2vec, along with a bilingual dictionary split into training and test sets. Such dictionaries are obtained from OPUS, including 5000 most frequent word pairs as the training set and 1500 randomly picked word pairs evenly distributed in 5 frequency bins. In terms of monolingual word emebeddings, the English training corpora consists of 2.8 billion words, including ukWaC, Wikipedia and BNC. The Italian training corpora includes 1.8 billion words for itWaC. German training corpora used SdeWac with 0.9 billion words, and Finnish training corpora used Finnish WMT 2016 dataset (Common Crawl). The Spanish word vectors are obtained by training WMT News Crawl 07 - 12, consisting of 386 million words.

In order to confirm our hypothesis, we extend this dataset to the eastern language family by adding Chinese-English pair. We train the word embeddings on a 1.5 billion word subset of the WMT 2018 Common Crawl corpora. Unlike Western language families, Chinese tokenisation needs a specific process to extract words from sentences. We adopt the solution from an open project: Jieba[3]. We train the word embeddings using the same configuration as Dinu et al. (2015). As for the dictionary, we take the English-Chinese dictionary provided by Lample et al. (2018), consisting of

[3] https://github.com/fxsjy/jieba

Table 1: A comparison of KCCA-based mapping with linear methods in terms of translation accuracy($\%$).

Method	**EN-IT**	**EN-DE**	**EN-ES**	**EN-FI**	**EN-ZH**
CCA	42.11	37.55	28.20	25.70	32.60
DCCA	43.53	43.13	34.86	25.28	45.34
Proposed	**48.4**	**50.13**	**38.86**	**37.43**	**52.56**

8728 training word pairs and 2230 test word pairs. However, we delete all of the out of vocabulary words in the dictionary, leaving 8239 training word pairs and 1964 test word pairs.

The evaluation of the benchmark is translation accuracy. Specifically, both word embeddings in test dictionaries are first projected using the learnt projection matrix. Then given a word in the test dictionary from source language, a retrieval approach is used to find the translation in the corresponding target language. The translation accuracy reflects the percentage of correct matches from source to target words. In the evaluation task, we adopt Cross-domain Similarity Local Scaling (CSLS) retrieval approach proposed by Lample et al. (2018).

4.2. Experiment setup

For our proposed KCCA-based cross-lingual word embedding model, we evaluated RBF and polynomial kernels and chose RBF. For RBF kernel, γ is tuned in the range [0, 1.5]. The weight for re-weighting is tuned in the range [0, 1] and the regularisation term κ is tuned in the range [0, 1].

For CCA, we tune the output dimension in the range from 150 to 300.

For DCCA, we employ the DCCA-based cross-lingual word embedding model proposed by Lu et al. (2015). However, their published model uses count-based word representations, so we tune a new parameter set. More concretely, we use two neural networks with linear units and ReLU activation function in hidden layers. The hidden layer size is tuned in $\{128, 256, 512, 1024, 2048, 4096\}$, the depth of the neural network is tune in $\{1, 2, 3, 4\}$. Parameters are tuned separately for each language. As for optimisation, we use stochastic gradient descent (SGD). The regularisation terms r_x and r_y are tuned in the range [1e-9, 1e-5]. All tuning was performed on the training set using 5-fold cross-validation.

4.3. Results

Table 1 shows that KCCA outperforms CCA, giving 6.29, 12.58, 10.66, 11.73 and 19.96 points improvement in English-Italian, English-German, English-Spanish, English-Finnish and English-Chinese datasets. Also, we show that after a proper fine-tuning process, DCCA-based non-linear mapping is better than CCA-based linear mapping approach. Judging from the improvement in the word translation task, the use of non-linear transformation outperforms linear transformation, confirming our hypothesis of the presence of non-linear relationships between some languages. It is also worth noting that KCCA outperforms DCCA on all datasets with an improvement of 4.3, 7, 4, 12.15 and 7.22 points. In our opinion, this is because deep neural networks can struggle to learn features from limited training data, and kernel-based methods do not suffer from data sparsity to the same extent.

Table 2 shows a comparison between our proposed approach with popular previous works, including supervised, semi-supervised and fully unsupervised scenarios. It is worth noting that the best result in a supervised setting is provided by Artetxe et al. (2018b), however, their retrieval approach is inverted softmax, which is a fair comparison with CSLS. Therefore, we reproduce their result using CSLS and also report it in Table 2. It can be seen that our proposed framework gets the best result among all supervised settings. Also, other than a close result in English-Italian dataset, our proposed approach achieved best results in all language pairs compared with unsupervised setting proposed by Artetxe et al. (2018a).

These results also leads to an interesting question, what kind of words are correctly translated when non-linear (kernel) approaches are used. We take EN-FI as an example. From Figure 1 we can see that CCA and KCCA correctly translate the same 320 English words (yellow points), KCCA is able to correctly translate 213 words that CCA fails to translate (orange points); however, CCA is able to translate only 46 words that KCCA fails on (green points). We believe the words that are correctly translated by KCCA but incorrectly translated by CCA have a higher possibility of exhibiting non-linear relationships; we denote the number of such words as N. We consider word pairs that are correctly translated by CCA to have a higher possibility of sharing a linear relationship and we denote the count of those word pairs as L. Then we define a ratio R:

$$R = \frac{N}{L} \tag{22}$$

as a measure to evaluate whether the relationship between two languages is more linear or non-linear; Table 3 provides the results.

Table 3 shows that Italian words have the highest possibility to share linear relationships with English. Most German and Spanish words could be mapped to English words with linear projections but there is also a considerable number of words could not be matched with such projections. Non-linear relationships have a huge impact on Finnish-English word pairs and Chinese-English word pairs.

In our opinion, this is because different languages have different grammars, which leads to different contexts for words with similar meanings in both languages, and the result confirms our hypothesis.

Figure 1: The translation result on EN-IT, EN-DE, EN-ES, EN-FI, EN-ZH test sets. The x axis represents the indices of English words. Yellow points indicates English words correctly translated by both CCA and KCCA. Orange points denote words correctly translated by KCCA only. Green points denote English words correcly translated by CCA only. Blue points denote incorrect translation by both CCA and KCCA. It can be observed that many more words are translated correctly by KCCA-only than CCA-only.

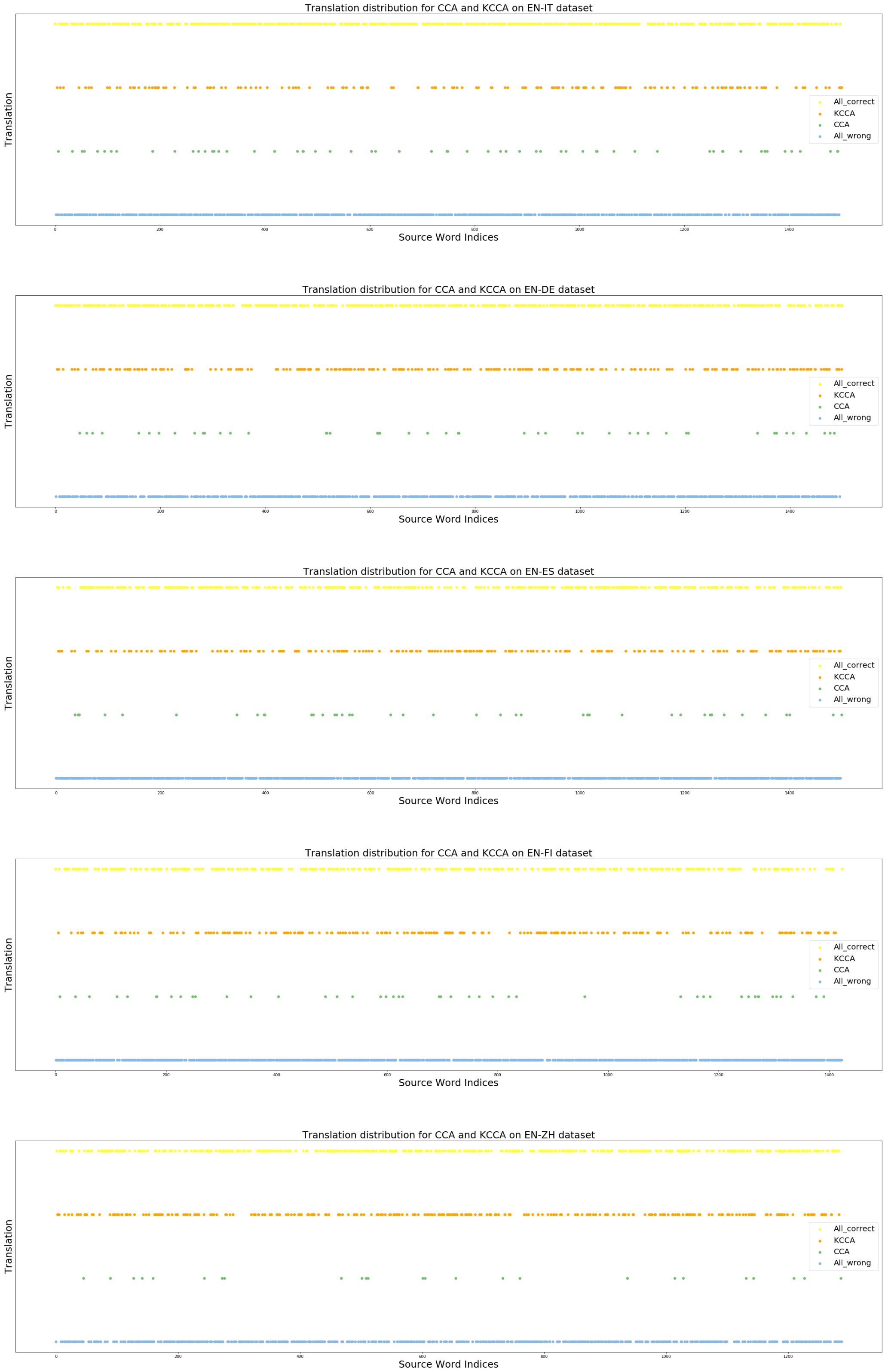

Table 2: A comparison of KCCA-based model with existing methods in terms of accuracy (%). All existing results are obtained from original papers, except results marked with *, which were produced by us using authors' original implementation.

Method	EN-IT	EN-DE	EN-ES	EN-FI	EN-ZH
Mikolov et al. (2013)	34.93	35.00	27.73	25.91	
Faruqui and Dyer (2014)	38.40	37.13	26.80	27.60	32.06*
Artetxe et al. (2016)	39.27	41.87	31.40	30.62	
Lu et al. (2015)	43.53	43.13	34.86	25.28	
Smith et al. (2019)	44.53	43.33	35.13	29.42	
Artetxe et al. (2018b)(Inverted softmax)	45.27	44.27	36.60	32.94	
Artetxe et al. (2018b)(CSLS)	47.33*	47.20*	38.20*	34.97*	49.20*
Artetxe et al. (2018a)	**48.53**	48.47	37.60	33.50	
Proposed	48.33	**50.13**	**38.86**	**37.43**	**52.56**

Table 3: A comparison of correct translations by CCA- and KCCA-based methods.

Method	EN-IT	EN-DE	EN-ES	EN-FI	EN-ZH
CCA only correct	60	46	41	46	25
Both correct	572	517	382	320	395
KCCA only correct	154	235	200	213	282
Ratio, R	24.4 %	41.7 %	47.3 %	58.2 %	**67.1%**

5. Conclusion

In this study, we question whether non-linear relationships exist between geometric arrangements of word vector representations of different languages and posit that non-linear mapping methods could produce better quality cross-lingual representations. Our experiments confirm our hypothesis and provide new state-of-art results.

6. Reference

Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., and Smith, N. A. (2016). Massively Multilingual Word Embeddings. *arXiv:1602.01925 [cs.CL].*

Artetxe, M., Labaka, G., and Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *The 2016 Conference on Empirical Methods in Natural Language Processing(EMNLP 2016)*, pages 2289–2294. Association for Computational Linguistics (ACL).

Artetxe, M., Labaka, G., and Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *55th Annual Meeting of the Association for Computational Linguistics (ACL2017).*

Artetxe, M., Labaka, G., and Agirre, E. (2018a). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *56th Annual Meeting of the Association for Computational Linguistics (ACL2018)*, 1:789–798.

Artetxe, M., Labaka, G., and Agirre, E. (2018b). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *32nd AAAI Conference on Artificial Intelligence (AAAI 2018).*

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018c). Unsupervised neural machine translation. In *6th International Conference on Learning Representations(ICLR 2018).*

Dinu, G., Lazaridou, A., and Baroni, M. (2015). Improving zero-shot learning by mitigating the hubness problem. In *3rd International Conference on Learning Representations (ICLR 2015).*

Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In *14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 462–471.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation*, 16(12):2639–2664, dec.

Lai, P. L. and Fyfe, C. (2000). Kernel and nonlinear canonical correlation analysis. *International journal of neural systems*, 10(5):365–377.

Lample, G., Conneau, A., Ranzato, M., Denoyer, L., and Jégou, H. (2018). Word translation without parallel data. In *6th International Conference on Learning Representations (ICLR 2018).*

Lu, A., Wang, W., Bansal, M., Gimpel, K., and Livescu, K. (2015). Deep multilingual correlation for improved word embeddings. *2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2015)*, pages 250–256.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013). Exploiting Similarities among Languages for Machine Translation. *arXiv:1309.4168 [cs.CL].*

Smith, S. L., Turban, D. H., Hamblin, S., and Hammerla, N. Y. (2019). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th Inter-*

national Conference on Learning Representations (ICLR 2017).

Tsai, C. T. and Roth, D. (2016). Cross-lingual wikification using multilingual embeddings. In *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2016)*.

Xing, C., Wang, D., Liu, C., and Lin, Y. (2015). Normalized word embedding and orthogonal transform for bilingual word translation. In *2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2015)*.