Kernel Least Squares Transformations for Cross-lingual Semantic Spaces

Adam Mištera[1][0009−0000−1019−9218] and Tomáš Brychcín[2][0000−0002−7442−0978]

[1] Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Czech Republic
amistera@kiv.zcu.cz

[2] NTIS – New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Czech Republic
brychcin@kiv.zcu.cz

Abstract. The rapid development in the field of natural language processing (NLP) and the increasing complexity of linguistic tasks demand the use of efficient and effective methods. Cross-lingual linear transformations between semantic spaces play a crucial role in this domain. However, compared to more advanced models such as transformers, linear transformations often fall short, especially in terms of accuracy. It is thus necessary to employ innovative approaches that not only enhance performance but also maintain low computational complexity.

In this study, we propose Kernel Least Squares (KLS) for linear transformation between semantic spaces. In our comprehensive analysis involving three intrinsic and two extrinsic experiments across six languages from three different language families and a comparative evaluation with nine different linear transformation methods, we demonstrate the superior performance of KLS. Our results show that the proposed method significantly improves word translation accuracy, thereby standing out as the most efficient method for transforming only the source semantic space.

Keywords: cross-lingual transformations · kernels · linear transformations · semantic spaces.

1 Introduction

Semantic spaces are based on the *Distributional Hypothesis* [18], which states that the meaning of a word is determined by its surroundings, so words that occur in similar contexts will also have similar meanings. Based on this hypothesis, several different semantic models were consequently developed [25,28,5].

The natural next step was the development of methods that would allow semantic spaces to be transformed between each other, or to create a unified space across several different languages. We can divide these methods for cross-lingual transformations into two basic groups, the supervised and unsupervised methods. Supervised methods are most often based on linear transformations

[26,2,13,22,6,19] and only need to build a dictionary containing a few thousand words [31]. The second group consists of unsupervised methods [11,21,3] that produce their own dictionary based on internal similarities in the given spaces. In this case, the dictionary is created only during the transformation and is of very high quality compared to the supervised approach.

Cross-lingual semantic spaces find application in many different NLP tasks. For instance, they can be used for sentiment analysis [27,1], document classification [20], or syntactic dependency parsing across languages [16]. Recently, more complex models based on the transformer architecture [30], such as BERT models [12,32], have been increasingly used for these tasks. However, experiments show that semantic spaces in certain cases can still achieve comparable quality at significantly lower computational cost [29].

In this paper, we present the *Kernel Least Squares* (KLS) method, a new approach proposed to improve the accuracy and efficiency of cross-lingual semantic transformations. Unlike traditional methods, the KLS method uses kernel-based techniques to capture nonlinear relationships in the data, which promises to significantly improve word translation accuracy and overall performance on a variety of language tasks.

The paper is organized as follows: Section 2 provides a detailed analysis of previous work in the field of cross-lingual transformations. In Section 3 we describe the kernels and our proposed transformation KLS. Section 4 presents the experimental setup, while Section 5 presents the measured results. Finally, we conclude in Section 6.

2 Cross-lingual Transformations between Semantic Spaces

A cross-lingual linear transformation between semantic spaces can be defined as:

$$\mathbf{Y} = \mathbf{T}^{x \to y}\mathbf{X}, \tag{1}$$

where matrices $\mathbf{X} \in \mathbb{R}^{m \times d}$ and $\mathbf{Y} \in \mathbb{R}^{m \times d}$ are constructed using the dictionary $\boldsymbol{D}^{x \to y}$ of word pairs (w^x, w^y), where x is the source language and y is the target language. Symbol m represents the size of the used dictionary, and d represents the dimension of the semantic space. In the following text, we present nine different cross-lingual transformations that were used for comparison in the experiments with our proposed transformation. They all differ in how they estimate the matrix $\mathbf{T}$.

The first of these transformations, the *Least Squares Transformation* [26], minimizes the total squared differences between paired word vectors in two languages aligned according to a bilingual dictionary. This approach offers an analytical solution through singular value decomposition (SVD). It uses a transformation matrix derived via the Moore-Penrose pseudoinverse, which provides a direct estimation for mapping between semantic spaces.

Orthogonal Transformation [2] extends the least squares approach by enforcing the orthogonality of the transformation matrix to preserve the angles

between the word vectors of the transformed semantic space. This method also uses SVD to compute the optimal transformation matrix.

The next method, *Canonical Correlation Analysis* [13] tries to maximize the correlation between projected vectors of two sets from different languages, finding optimal basis vectors for each set. This method uses canonical directions to project words into a shared semantic space, enhancing multilingual semantic performance.

Ranking Transformation [22] employs a max-margin hinge loss to reduce hubness in high-dimensional spaces, optimizing the alignment of words across languages by adjusting their ranks. The goal is to ensure the correct alignment of words across languages by optimizing their relative ranks within the semantic space.

Orthogonal Ranking Transformation [6] adds orthogonal constraints to the Ranking Transformation to address asymmetry problem in cross-lingual mappings. It uses two max-margin loss functions to optimize transformation matrix towards being *nearly orthogonal*, thus preserving the monolingual performance.

Geometry-aware Multilingual Mapping [19] aligns semantic spaces using orthogonal transformations and *Mahalanobis* metrics allowing for more efficient learning of similarity measures. This method optimizes both the source and target semantic spaces to achieve more accurate semantic alignment.

Vector Mapping [3] is an unsupervised method that iteratively refines the transformation matrix without the need for labeled parallel data, which is ideal for resource-limited languages. It starts by aligning semantic spaces and refines the mapping using a self-learning algorithm focusing on the internal structures of the semantic spaces.

Multilingual Unsupervised and Supervised Embeddings [11] is an unsupervised method that uses adversarial training to align semantic spaces without the need for bilingual dictionaries. It improves the alignment quality by using a discriminator that tries to distinguish between transformed and real word vectors of the target semantic space.

Kernel Canonical Correlation Analysis [4] is method based on the previously mentioned canonical correlation analysis, but enhances it by using kernels. By using kernels, the nonlinear relationship between two languages can be captured using a linear transformation. The obtained transformation matrix can thus be used in the same way as in the previous transformations.

3 Proposed Transformation

In the proposed *Kernel Least Squares Transformation* method, we build on the first proposed transformation, the *Least Squares Transformation* method, but greatly improve it by using *kernels*. The use of kernels allows us to significantly enhance the accuracy of the transformations by facilitating the capture of complex, high-dimensional relationships in the data. Since kernels are an essential part of our transformation, we will briefly introduce them in more detail in the following Section 3.1.

3.1 Kernel Function

The use of a *kernel* or *kernel function* $\kappa : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ in the proposed transformation allows us to avoid the explicit mapping that is necessary for linear learning algorithms to model a nonlinear function. We define the kernel between any two data inputs a_i and a_j from the space $\mathcal{S} = \{a_1, \ldots, a_n\}$ as:

$$\kappa(a_i, a_j) = \langle \varphi(a_i), \varphi(a_j) \rangle, \tag{2}$$

where $\varphi : \mathcal{S} \to \mathcal{F}$ represents a mapping to a feature space $\mathcal{F}$, and the function $\langle \cdot, \cdot \rangle$ denotes a generalized inner product in the feature space, that can be used for both vectors and matrices. This transformation of inputs a_i and a_j into a higher-dimensional space is essential for identifying latent structures in semantic spaces and enhancing transformations.

To further explain the use of kernels in the transformation, it is useful to introduce the concept of *kernel matrix* or *Gram matrix*. This $n \times n$ matrix, denoted as $\mathbf{K}$, is created by evaluating the chosen kernel function for each pair of data points in a given data set. The entries of the matrix are given as:

$$\mathbf{K}_{ij} = \langle \varphi(a_i), \varphi(a_j) \rangle = \kappa(a_i, a_j). \tag{3}$$

The kernel matrix is symmetric since $\mathbf{K}_{ij} = \mathbf{K}_{ji}$, and contains the pairwise comparisons within the space.

Table 1. Kernel types.

Kernel Function	Kernel Formula
LK	$\kappa(a_i, a_j) = a_i^T a_j$
PK	$\kappa(a_i, a_j) = (a_i^T a_j + c)^d$
RBF	$\kappa(a_i, a_j) = \exp\left(-\gamma \|a_i - a_j\|_2^2\right)$
CSK	$\kappa(a_i, a_j) = a_i^T a_j / (\|a_i\|_2 \|a_j\|_2)$

Several commonly used types of kernels are defined in the Euclidean space $\mathbb{R}^d$. Choosing the right kernel type is crucial for the final performance of the transformation, as different kernels can significantly affect the resulting quality of transformed semantic space and thus the results of individual experiments. In our experiments, we employ four different types of kernels, namely, *Linear Kernel* (LK), *Polynomial Kernel* (PK), *Radial Basis Function Kernel* (RBF), and *Cosine Similarity Kernel* (CSK). The formulas for their calculation are given in Table 1. Each of these kernels offers distinct advantages and is key to the success of the semantic space transformation. Most of these kernels, especially RBF, are commonly encountered in the context of SVM algorithms, which make extensive use of them to find nonlinear decision boundaries [10]. However, in the field of NLP we often see polynomial kernels of lower degrees [10,14].

3.2 Kernel Least Squares Transformation

The optimization criterion is defined by Equation 4, where $\mathbf{K}$ denotes the kernel matrix computed from the input data $\mathbf{X}$, encapsulating the similarities between data points. The symbol $\hat{\mathbf{T}}^{x \to y}$ indicates the optimal transformation matrix.

$$\hat{\mathbf{T}}^{x \to y} = \underset{\mathbf{T}^{x \to y}}{\arg\min} \left\| \mathbf{Y} - \mathbf{K}\mathbf{T}^{x \to y} \right\|_2^2. \tag{4}$$

If we take the derivative of the equation with respect to $\hat{\mathbf{T}}^{x \to y}$, then the analytic solution is given as $\hat{\mathbf{T}}^{x \to y} = \mathbf{K}^{-1}\mathbf{Y}$. However, in practice, to increase stability and ensure that the $\mathbf{K}$ matrix is invertible, we add the regularization term λ. The final equation with this term is then given by:

$$\hat{\mathbf{T}}^{x \to y} = (\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{Y}. \tag{5}$$

Here $\lambda > 0$ is a scalar value that adjusts the regularization strength, and $\mathbf{I}$ represents the identity matrix.

Subsequently, we can transform arbitrary word vector $\mathbf{x}$ from source semantic space to target semantic space with the following equation

$$\hat{y} = \kappa(\mathbf{x}, \mathbf{X})\hat{\mathbf{T}}^{x \to y}. \tag{6}$$

Essentially, we create a new kernel matrix for the transformed vector $\mathbf{x}$ that contains the calculated similarities of this vector to all vectors from the input data $\mathbf{X}$ in the given feature space. Therefore, this equation can be used to easily transform each word vector from the source to the target semantic space.

The whole process of the proposed transformation method then involves several key steps. First, we construct a kernel matrix $\mathbf{K}$ with the chosen kernel function $\kappa(\mathbf{x}_i, \mathbf{x}_j)$, which we evaluate for each pair of vectors $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$. We then calculate the optimal transformation matrix $\hat{\mathbf{T}}^{x \to y}$ that achieves the smallest values of the loss function as given in Equation 4. In the last step, it is necessary to compute a kernel $\kappa(\mathbf{x}, \mathbf{X})$ between all $\mathbf{x}$ from the source semantic space $\mathcal{S}^x$ and matrix $\mathbf{X}$, enabling the transformation of the word vectors according to Equation 6. After these steps, the transformation is complete and the semantic space can then be used in any subsequent downstream tasks.

3.3 Preprocessing and Postprocessing

A necessary step to ensure optimal transformation quality is the appropriate configuration and application of preprocessing and postprocessing techniques to the semantic spaces. Both semantic spaces are therefore first column-wise centered, followed by unit vector normalization, as shown in [2]. Normalizing the semantic spaces ensures that the vectors contribute equally to the transformation and their length does not affect the mapping.

Our research showed that applying the above steps to the source semantic space after the transformation has been performed can further improve the measured results. For this reason, in Section 5, we present the results of methods

with this setup, i.e. postprocessing. In the following Section 4, we present the experiments used to evaluate the quality of the proposed transformation and the measured results.

4 Experiments

In total, we conducted five different experiments, three intrinsic ones, namely *Word Translation* (WT), *Cross-lingual Word Analogies* (WA), and *Cross-lingual Word Similarities* (WS), and two extrinsic ones, *Topic Classification* (TC), and *Sentiment Analysis* (SA). The individual experiments are described in more detail in Section 4.2. All experiments were performed in both directions, i.e. from the source semantic space to the target space and vice versa. This allowed us to more accurately assess the resulting quality of individual transformations, since some transformations may perform very well in the first direction and fail in the opposite direction due to the asymmetry problem [6].

Nine additional linear transformations have been tested, including seven supervised methods, namely *Least Squares Transform* (LS), *Orthogonal Transform* (OT), *Canonical Correlation Analysis* (CCA), *Ranking Transform* (RT), *Orthogonal Ranking Transformation* (ORT), *Geometry-aware Multilingual Mapping* (GEOMM), and *Kernel Canonical Correlation Analysis* (KCAA) and two unsupervised methods, namely *Vector Mapping* (VM) and *Multilingual Unsupervised and Supervised Embeddings* (MUSE). The evaluation is conducted on six languages from different language families, namely Czech (Cs), Croatian (Hr), English (En), German (De), Italian (It), and Spanish (Es).

4.1 Experimental Setup

To create the transformation matrices $\mathbf{X} \in \mathbb{R}^{m \times d}$ and $\mathbf{Y} \in \mathbb{R}^{m \times d}$, where $m = 20,000$ as recommended in [6], we first translated the top $50,000$ words from the source language into the target language using *Google Translate*. We then divided the created bilingual dictionary into train and test subsets, reserving the latter for later use in the word translation task. For all tested languages we employed semantic spaces with dimension size $d = 300$ pre-trained on a corpus combined from *Common Crawl* and *Wikipedia* [15].

All transformations were performed with the recommended settings and with the previously mentioned dictionary size. However, only the source semantic space was transformed for the experiments, while the target space remained unchanged to ensure a fair comparison, except for the MUSE and KCCA methods, which do not support this constraint. The RT and ORT methods were configured to use five negative samples, with the parameter $\gamma = 0.275$. The VM and MUSE transformations were run in completely unsupervised mode without a provided train dictionary. Preprocessing was applied to all methods before transformation, and the process was repeated after the transformation was completed.

4.2 Evaluation Metrics

Word Translation In this experiment, words from the source semantic space are transformed into the target semantic space using a transformation matrix, and the success of finding correct translations is tested by checking whether the k-nearest neighbors of the transformed word vector contain an accurate translation. Two different settings of the number of neighbors were used to evaluate the accuracy of translations, specifically $k = 1$ and $k = 5$. The test part of the dictionary was used to evaluate results. This experiment was performed for all language pairs.

Cross-lingual Word Analogies This experiment serves for evaluating cross-lingual word analogies using a dataset divided into semantic and syntactic categories [7]. The experiment involves predicting a word that completes an analogy based on relationships between word pairs across languages. To find the target word that best matches the analogy, vector operations are used. As in the previous case, this experiment was conducted in two settings ($k = 1$ and $k = 5$). This experiment was performed for all language pairs.

Cross-lingual Word Similarities In this intrinsic experiment, we evaluate cross-lingual word similarities using three datasets with predefined word pair similarities in multiple languages, namely *RG-65* [9], *SemEval2017* [8], and *WS353* [9]. Each dataset uses a scale to measure the similarity between word pairs, with scores provided by human annotators. The experiment evaluates the quality of transformations by comparing the cosine similarity of the word vectors with provided scores using Pearson's correlation coefficient, where a higher correlation means a more accurate transformation.

Topic Classification First extrinsic evaluation involves topic classification using the RCV2 Reuters dataset [23], where documents are categorized into four predefined categories in *English*, *German*, and *Spanish*. For classification evaluation, we train two different models based on *Convolutional Neural Networks* (CNN) and *Long Short-Term Memory* (LSTM) architectures, which contain significantly fewer parameters than transformers. The experiment tests the classification accuracy using *F-measure*. The experiment was divided into four subtasks. Firstly, the CNN was trained on the transformed source semantic space, and the classification was evaluated on the unmodified target space. Then, the experiment was repeated with training on the target space and evaluation on the source space. The same experiments were also performed for the LSTM.

Sentiment Analysis Finally, a sentiment analysis experiment for movie reviews was conducted using two datasets for *Czech* and *English*, divided into positive and negative sentiments. The datasets, CSFD [17] and IMDB [24], involve $90,000$ and $50,000$ reviews respectively. Similar to the previous classification task, this experiment employs CNN and LSTM models, with performance evaluated using the *F-measure*. As in the previous case, this experiment was divided into four subtasks.

5 Results

Table 2 shows the complete results measured in our experiments for all tested methods for transforming semantic spaces. The numbers in the columns correspond to the measured results for the individual experiments discussed in Section 4.2. As shown in the table, the proposed method with the *polynomial kernel* achieved the best result among all the transformations tested with an average score across all task equal to 0.655. The proposed method particularly excels at the word translation task, where it achieved the highest value. With a value of 0.452, it outperforms the second-best method by more than three percent, representing a significant improvement for this category. The second place was also taken by the proposed method, but this time with the *radial basis function kernel*, achieving an average score 0.645, which is comparable to the result of the ORT method. The KCCA transformation, which also uses kernels, performed especially well in the topic classification and sentiment analysis tasks, where it achieved the best scores.

Table 2. Overall results of all experiments. For each experiment, we show the average across all languages tested in the experiment and across both directions of transformation. The last column, denoted as AVG, contains the average score across all experiments performed in both directions.

	WT	WA	WS	TC	SA	AVG
LS	0.393	0.559	0.685	0.753	0.786	0.635
OT	0.382	0.569	0.694	0.747	0.784	0.635
CCA	0.385	**0.570**	0.691	0.752	0.786	0.636
RT	0.363	0.547	0.691	0.716	0.766	0.617
ORT	0.419	0.561	**0.727**	0.729	0.786	0.644
GEOMM	0.413	0.530	0.732	0.725	0.778	0.636
VM	0.336	0.537	0.698	0.755	0.784	0.622
MUSE	0.281	0.489	0.675	0.748	0.747	0.588
KCCA	0.329	0.517	0.669	**0.798**	**0.811**	0.625
KLS+LK	0.393	0.559	0.685	0.756	0.785	0.636
KLS+CSK	0.393	0.559	0.685	0.752	0.783	0.634
KLS+RBF	0.418	0.567	0.706	0.746	0.786	0.645
KLS+PK	**0.452**	0.563	0.719	0.748	0.793	**0.655**

The first two kernels, specifically the *linear kernel* and the *cosine similarity kernel*, achieved comparable results to the ordinary least squares method, especially in the first three categories. The results suggests that these kernels therefore do not introduce any additional information to the transformation that would improve it in the result. This not only shows that the linear kernel and the cosine similarity kernel are completely equivalent when using unit normalization of semantic space vectors, but also highlights the importance of an appropriate choice of preprocessing and postprocessing techniques for semantic spaces.

The *radial basis function kernel* performed overall very well, improving on the previous two kernels in all measured categories. It also performed the best of all the kernels tested in the cross-lingual word analogy category. Despite its popularity and frequent use, however, it has not achieved the best results in these experiments overall. This may be caused by the overly complex nature of the kernel, as the feature space of this kernel has infinitely many dimensions, which may reduce the quality of the resulting transformation.

The best performing kernel was the *polynomial kernel*. We conducted several experiments with different settings of the polynomial degree for this kernel, among which the kernel with polynomial of degree $d = 6$ performed the best. The final scores of the measured experiments gradually decreased with increasing and decreasing polynomial degree, respectively.

Table 3. Average results of the KLS method over all experiments for all language pairs.

	Cs	**De**	**En**	**Es**	**Hr**	**It**
Cs	1.000	0.478	0.664	0.449	0.424	0.462
De	0.504	1.000	0.699	0.612	0.440	0.535
En	0.681	0.683	1.000	0.696	0.554	0.645
Es	0.459	0.574	0.689	1.000	0.415	0.614
Hr	0.451	0.411	0.510	0.388	1.000	0.402
It	0.484	0.515	0.645	0.623	0.437	1.000

The Table 3 shows the results for each language pair for the KLS method. In the table we can see that transformations to or from the English semantic space yield the best results. This is an expected behavior, since English is the most resource-rich language and thus its semantic spaces are generally of high quality. To improve the results of individual tasks, it is therefore worth transforming the semantic space into such a space.

6 Conclusion

In this paper, we introduced a new transformation method, *Kernel Least Squares*, which uses kernels to improve the quality of the transformation and the resulting semantic space. Our findings show that this newly proposed method outperforms existing transformations that focus only on modifying the source semantic space, on average across all languages and experiments. It also completely dominates in the word translation task compared to all tested transformations. At the same time, *Kernel Canonical Correlation Analysis* achieved the best scores for the topic classification and sentiment analysis tasks. Thus, it can be concluded that kernel transforms are definitely a promising direction.

The main advantage of the proposed method is significantly lower computational complexity compared to more advanced models based on transformer

architecture. In addition, the newly proposed method provides opportunities for further improvement by designing custom kernels that can further optimize the resulting transformation.

section*Acknowledgments This work has been partly supported by grant No. SGS-2022-016 Advanced methods of data processing and analysis. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

1. Abdalla, M., Hirst, G.: Cross-lingual sentiment analysis without (good) translation. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 506–515. Asian Federation of Natural Language Processing, Taipei, Taiwan (Nov 2017), https://aclanthology.org/I17-1051
2. Artetxe, M., Labaka, G., Agirre, E.: Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2289–2294. Association for Computational Linguistics, Austin, Texas (November 2016), https://aclweb.org/anthology/D16-1250
3. Artetxe, M., Labaka, G., Agirre, E.: A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 789–798. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1073, https://aclanthology.org/P18-1073
4. Bai, X., Cao, H., Zhao, T.: Improving vector space word representations via kernel canonical correlation analysis. ACM Trans. Asian Low-Resour. Lang. Inf. Process. **17**(4) (jul 2018). https://doi.org/10.1145/3197566, https://doi.org/10.1145/3197566
5. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017), https://transacl.org/ojs/index.php/tacl/article/view/999
6. Brychcín, T.: Linear transformations for cross-lingual semantic textual similarity. Knowledge-Based Systems **187**, 104819 (2020)
7. Brychcín, T., Taylor, S., Svoboda, L.: Cross-lingual word analogies using linear transformations between semantic spaces. Expert Systems with Applications **135**, 287–295 (2019). https://doi.org/https://doi.org/10.1016/j.eswa.2019.06.021, https://www.sciencedirect.com/science/article/pii/S0957417419304191
8. Camacho-Collados, J., Pilehvar, M.T., Collier, N., Navigli, R.: Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 15–26. Association for Computational Linguistics, Vancouver, Canada (August 2017), http://www.aclweb.org/anthology/S17-2002
9. Camacho-Collados, J., Pilehvar, M.T., Navigli, R.: A framework for the construction of monolingual and cross-lingual word similarity datasets. In: Proceedings of ACL (2). pp. 1–7 (2015)

10. Chang, Y.W., Hsieh, C.J., Chang, K.W., Ringgaard, M., Lin, C.J.: Training and testing low-degree polynomial data mappings via linear svm. Journal of Machine Learning Research **11**(48), 1471–1490 (2010), http://jmlr.org/papers/v11/chang10a.html
11. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. arXiv preprint arXiv:1710.04087 (2017)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
13. Faruqui, M., Dyer, C.: Improving vector space word representations using multilingual correlation. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 462–471. Association for Computational Linguistics, Gothenburg, Sweden (April 2014), http://www.aclweb.org/anthology/E14-1049
14. Goldberg, Y., Elhadad, M.: splitSVM: Fast, space-efficient, non-heuristic, polynomial kernel computation for NLP applications. In: Moore, J.D., Teufel, S., Allan, J., Furui, S. (eds.) Proceedings of ACL-08: HLT, Short Papers. pp. 237–240. Association for Computational Linguistics, Columbus, Ohio (Jun 2008), https://aclanthology.org/P08-2060
15. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018) (2018)
16. Guo, J., Che, W., Yarowsky, D., Wang, H., Liu, T.: 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). pp. 1234–1244. Association for Computational Linguistics, Beijing, China (July 2015), http://www.aclweb.org/anthology/P15-1119
17. Habernal, I., Ptáček, T., Steinberger, J.: Sentiment analysis in Czech social media using supervised machine learning. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 65–74. Association for Computational Linguistics, Atlanta, Georgia (Jun 2013), https://aclanthology.org/W13-1609
18. Harris, Z.: Distributional structure. Word **10**(23), 146–162 (1954)
19. Jawanpuria, P., Balgovind, A., Kunchukuttan, A., Mishra, B.: Learning Multilingual Word Embeddings in Latent Metric Space: A Geometric Approach. Transactions of the Association for Computational Linguistics **7**, 107–120 (04 2019). https://doi.org/10.1162/tacl_a_00257, https://doi.org/10.1162/tacl_a_00257
20. Klementiev, A., Titov, I., Bhattarai, B.: Inducing crosslingual distributed representations of words. In: Proceedings of COLING 2012. pp. 1459–1474. The COLING 2012 Organizing Committee, Mumbai, India (December 2012), http://www.aclweb.org/anthology/C12-1089
21. Lample, G., Conneau, A., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=H196sainb

22. Lazaridou, A., Dinu, G., Baroni, M.: Hubness and pollution: Delving into cross-space mapping for zero-shot learning. 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). pp. 270–280. Association for Computational Linguistics, Beijing, China (July 2015), http://www.aclweb.org/anthology/P15-1027
23. Lewis, D.D., Yang, Y., Russell-Rose, T., Li, F.: Rcv1: A new benchmark collection for text categorization research. Journal of machine learning research **5**(Apr), 361–397 (2004)
24. Maas, A., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. pp. 142–150 (2011)
25. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013), http://arxiv.org/abs/1301.3781
26. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. CoRR **abs/1309.4168** (2013), http://arxiv.org/abs/1309.4168
27. Mogadala, A., Rettinger, A.: Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 692–702. Association for Computational Linguistics, San Diego, California (June 2016), http://www.aclweb.org/anthology/N16-1083
28. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (October 2014), http://www.aclweb.org/anthology/D14-1162
29. Přibáň, P., Šmíd, J., Mištera, A., Král, P.: Linear transformations for cross-lingual sentiment analysis. In: Text, Speech, and Dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic, September 6–9, 2022, Proceedings. pp. 125–137. Springer (2022)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
31. Vulić, I., Korhonen, A.: On the role of seed lexicons in learning bilingual word embeddings. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 247–257. Association for Computational Linguistics, Berlin, Germany (August 2016), http://www.aclweb.org/anthology/P16-1024
32. Zhuang, L., Wayne, L., Ya, S., Jun, Z.: A robustly optimized BERT pre-training approach with post-training. In: Proceedings of the 20th Chinese National Conference on Computational Linguistics. pp. 1218–1227. Chinese Information Processing Society of China, Huhhot, China (Aug 2021), https://aclanthology.org/2021.ccl-1.108