Cross-lingual Neural Vector Conceptualization

Lisa Raithel*[1,2] and Robert Schwarzenberg*[1]

[1] German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
[2] Giance Technologies GmbH, Berlin, Germany
lisa.raithel@giance.ai

Abstract. Recently, Neural Vector Conceptualization (NVC) was proposed as a means to interpret samples from a word vector space. For NVC, a neural model activates higher order concepts it recognizes in a word vector instance. To this end, the model first needs to be trained with a sufficiently large instance-to-concept ground truth, which only exists for a few languages. In this work, we tackle this lack of resources with word vector space alignment techniques: We train the NVC model on a high resource language and test it with vectors from an aligned word vector space of another language, without retraining or fine-tuning. A quantitative and qualitative analysis shows that the NVC model indeed activates meaningful concepts for unseen vectors from the aligned vector space. NVC thus becomes available for low resource languages for which no appropriate concept ground truth exists.

Keywords: Interpretability · Explainability · Word Vector Space.

1 Introduction

Neural Vector Conceptualization [14] is an interpretability method that allows to illuminate continuous, distributed word vector spaces with higher order concepts. To this end, NVC maps samples from a word vector space into a concept space, with the help of a neural network. This neural mapping is learned in advance, in a supervised manner, using a pre-trained embedding space to draw training instances from and an instance-to-concept graph to retrieve appropriate target concepts. For example, we expect the instance "Confucius" to activate the concept "philosopher." Unfortunately, the need for a sufficiently large concept graph impedes the deployment of NVC in low resource languages.

It has been shown, however, that continuous word vector spaces of different languages share structural properties, which is a feature that word vector alignment methods make use of [12,4]. A word vector alignment aims to minimize the distance between word vectors of parallel words in a source and a target language. To this end, a transformation from the source to the target space is learned such that $W \in R^{d \times d}$ minimizes a distance δ according to

$$\frac{1}{n} \sum_{i=1}^{n} \delta(W x_i, y_i) \tag{1}$$

*Shared first authorship.

where $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ is a seed dictionary of parallel d-dimensional word vectors from the source and target language X and Y, respectively [9]. In this work, we train the NVC model with a high resource language Y and test it with aligned vectors from language X, aiming to make the methods accessible for low resource languages through vector space alignments.

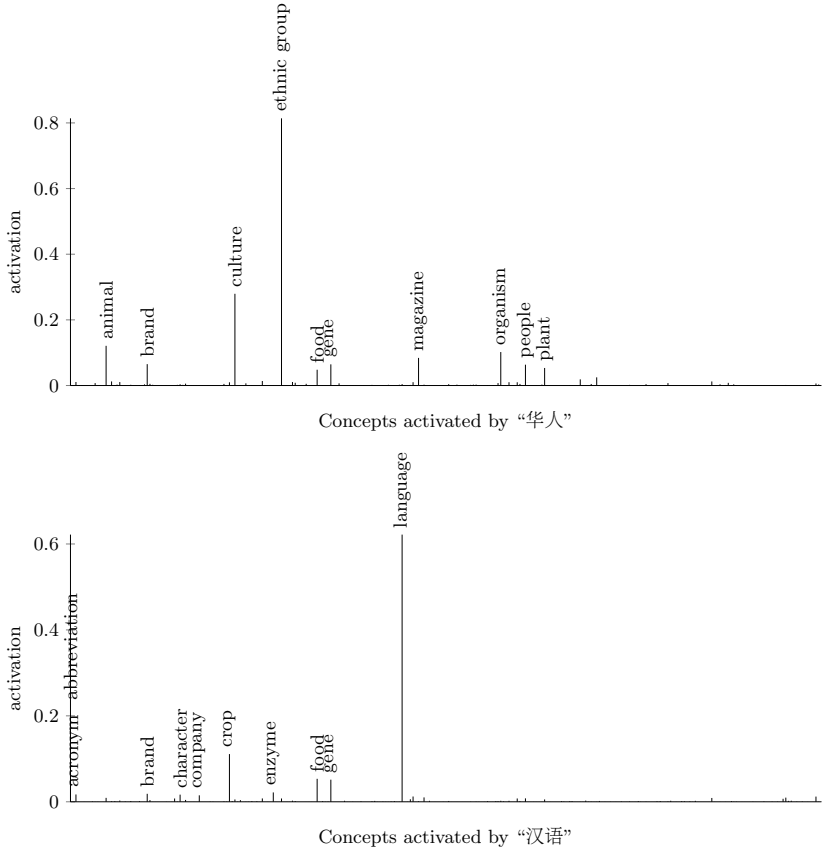

Fig. 1. Neural Vector Conceptualizations. Top: NVC of the word vector of the Chinese translation for "Chinese people." Bottom: NVC of the word vector of the Chinese translation for "Chinese language." Both word vectors were retrieved from a fastText word vector space that has been aligned with the English fastText word vector space the NVC model was trained on. The model was not retrained or fine-tuned on the aligned vector space. Consequently, both vectors were not seen during training.

2 Methods and Experiments

Following [14], for NVC, we trained a neural multi-layer feed-forward net to recognize concepts in word vectors. We drew the target concepts for each word vector training instance from the Microsoft Concept Graph (MCG) [15], using the same preprocessing steps and preprocessing hyperparameters as the authors of NVC. Preprocessing involved filtering concepts in the MCG that have a sufficient number of instances in the word vector space.

After preprocessing, 32768 instances and 275 associated concepts remained for training the NVC model (+ 8192 for validating, + 4553 for testing). We optimized with the Adam optimizer [10] with a learning rate of 0.001 and stopped training early with a patience of four epochs. The best model was determined in a five-fold cross validation procedure.

Our choice for a class membership threshold deviated from the experiments of [14]. We did not use a strict class membership threshold of 0.5 but instead optimized the threshold on the dev split after training. This step was motivated by the observation that for some word vectors the top k concept activations were meaningful but below the 0.5 threshold.

Contrary to [14], instead of with word2vec [11], we trained with English fastText vectors [1], since for these, alignments already exist in more than 40 languages.[3] For the alignments, the Relaxed Cross-Domain Similarity Scaling (RCSLS) method, which mitigates issues with high degree neighborhoods in the word vector space, was used. Here, we omit details and instead refer the interested reader to the original paper [9].

After training, the best model was used to conceptualize Chinese word vectors from an aligned fastText space. It is noteworthy that the model had never seen a vector from that space during training. To statistically validate that NVC works across language boundaries for aligned word vector spaces, without retraining, we determined the concept classification performance of the model on the unseen Chinese vectors. To this end, we treated the target concepts of parallel English words as the ground truth.

Parallel English and Chinese tokens were retrieved[4] from the 10k most frequent tokens in the Sinica Corpus [3] and then intersected with the above mentioned 45513 instances. The intersection contained 1,288 instances. In the next section, we report on the F scores our model achieved in the English and Chinese task and we also present NVCs of selected word vectors. Our experiments are publicly available under `https://github.com/dfki-nlp/cross-nvc`.

3 Results & Discussion

The results of our classification experiments on the English test set are summarized in Table 1. We optimized the class membership threshold on the dev

[3] Word vectors retrieved from `https://fasttext.cc/docs/en/aligned-vectors.html` on 2019/07/16.

[4] Retrieved from `https://en.wiktionary.org/wiki/Appendix:Mandarin_Frequency_lists` on 2019/07/30.

set, which yielded a threshold of 0.18. A stricter threshold of 0.5 resulted in a weighted F score of only 0.29 which is closer to the F score of 0.22, reported by [14].

[14], however, trained with over 600 concepts while we trained with only 275 concepts because the intersection of word vector vocabulary and MCG was smaller in our case. We can assume that a lower-dimensional concept space facilitates the task which should be one reason why our model performs better, in terms of F measure.

Table 1. Results on a held-out test set of 4553 English instances with 275 concepts. There are on average 1.14 concepts per instance.

	weighted	macro	micro
precision	0.365	0.327	0.393
recall	0.375	0.334	0.375
F1	0.351	0.309	0.384

The low, optimized threshold supports the authors' suspicion that a conservative 0.5 threshold is too strict for the task since apparently, the top k activations for some vectors contain meaningful concepts below this threshold. Furthermore, we note that our F score is considerably above chance. We take this as evidence that NVC works with fastText vectors, too.

Table 2. Results on test set of 1,288 Chinese word vectors with 275 concepts.

	weighted	macro	micro
precision	0.132	0.048	0.092
recall	0.180	0.060	0.180
F1	0.096	0.038	0.122

To determine whether NVC also works across language boundaries if aligned word vector spaces are used, we repeated the concept classification experiment with aligned Chinese word vectors that we retrieved as described above. The performance of our model, which we did not retrain, is summarized in Table 2.

We observe that unsurprisingly, the model performs worse when tested with Chinese word vectors but that its performance is again above chance with a considerable margin. We take this as evidence that NVC indeed works across language boundaries. For the loss in accuracy several reasons can be cited:

Firstly, errors and ambiguities in the dictionary we used to retrieve aligned word vectors worsen performance. In fact, inaccuracies in this step lead to the retrieval of unaligned vectors which should not activate the concepts of their supposed parallel counter parts in the English language.

Secondly, the alignment of the two vector spaces we drew samples from did not happen without loss. We can thus expect noise in aligned word vectors even if

their meanings in the corpora they origin from were culturally and semantically identical. Our model never learned to filter this noise.

Lastly, even if we retrieved the most accurate translation for an instance from our dictionary, we cannot expect it to have the exact same meaning, as assumed above. Some concepts will be lost in translation, others will be activated.

Consider, for example, a culture in which people consume soups for breakfast but not for supper and another one in which it is vice versa. This could be reflected in the NVCs of the instance `soup`: While the one vector may activate `breakfast`, it may be deactivated (correctly) in the other, whereas `supper` may become activated (correctly). Cultural differences of course worsen performance. In the example above, concepts correctly activated by the translation may not be reflected in the original instance-to-concept knowledge base. Nevertheless, arguably, this could be considered a feature rather than a bug.

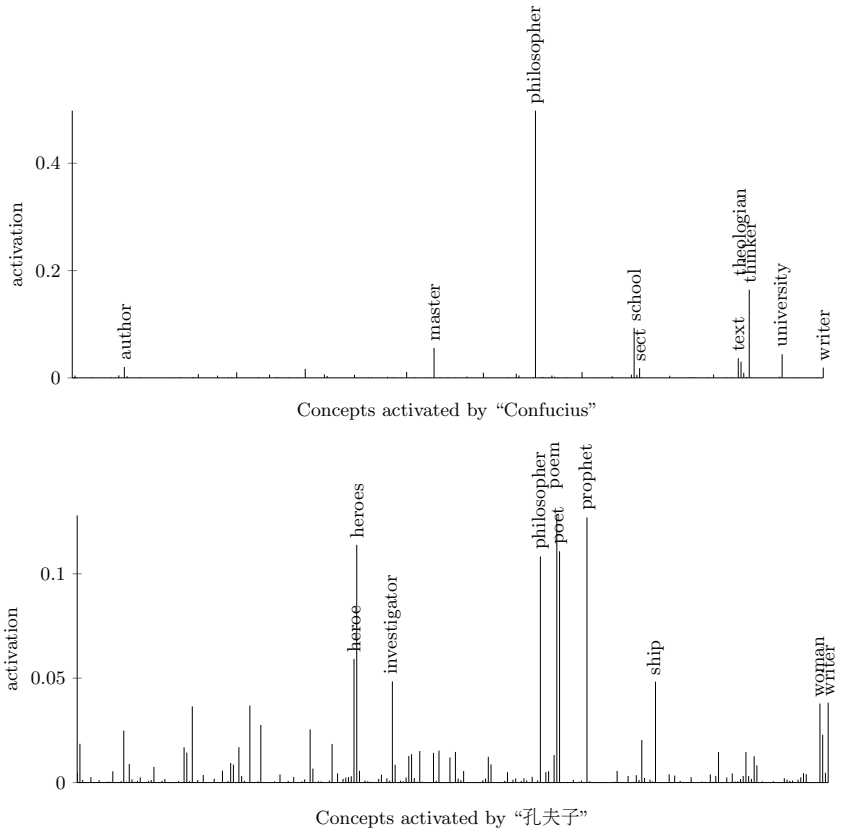

Fig. 2. Neural Vector Conceptualizations. Top: NVC of the English word vector (unseen during training) for "Confucius" from the fastText space the NVC model was trained on. Bottom: NVC of the word vector of the Chinese translation from an aligned word vector space from which no sample was drawn during training.

We now conduct qualitative analysis. Figs. 1 and 2 show selected NVCs of Chinese and English word vectors. In Fig. 1 we present the NVCs of the Chinese word vectors for the terms `Chinese people` and `Chinese language`, both of which were sampled from the aligned Chinese fastText space, unseen during training. `ethnic group` and `people` are activated in the former NVC, while in the latter `language` is the top activation. We believe these to be meaningful concepts which supports our claim that NVC produces worthwhile activation profiles across language boundaries, if alignment techniques are used.

Cultural differences in word vectors might be reflected in Fig. 2. The English word vector for `Confucius` strongly activates `philosopher` while the top concept activated by the aligned vector of the Chinese translation is `poem`. We also find `poet` and `writer` under the top concepts activated by the Chinese vector with no such strong artistic connotation in the English NVC. One reason for these differences in the NVCs may be that Chinese corpora emphasize the artistic side of Confucius stronger than western corpora.

4 Related Work

In addition to related work already discussed in [14], we would like to point out the recent work by [6]. Similar to us and [14], they also use concepts for interpretation, which they automatically extract from images.

Regarding the cross-lingual capacities of NVC which we focus on here, there are several other branches of related work on multilingualism worth mentioning. [7], for instance, provide an extensive overview of cross-lingual embedding techniques and propose several measures on how to evaluate them. They argue that projection-based cross-lingual embeddings, like those of [9], are usually only evaluated on bilingual lexicon induction and offer additional downstream evaluation tasks, for example cross-lingual information retrieval. NVC can be regarded as yet another evaluation downstream task.

[8] evaluate cross-lingual word vectors on an ontology alignment task, aiming to identify overlapping concepts in multi-lingual ontologies. Cross-lingual NVC might be used for an alternative embedding evaluation and may be also useful for ontology alignment.

We additionally can relate our method to research that is concerned with biases in word embeddings. There exists a plethora of work concerned with identifying and removing biases in embeddings, see for example [2,5,13].

5 Conclusion & Future Directions

In this work, we made Neural Vector Conceptualization available to low resource languages, for which no sufficiently large instance-to-concept ground truth exists. To this end, we trained the NVC model with a word vector space of a high resource language (English) and tested it with an aligned vector space of another language (Chinese).

Quantitative experimental results strongly suggest that NVC indeed works across language boundaries, if aligned vector spaces are used. A qualitative analysis revealed that meaningful concepts were activated for unseen vectors from the aligned vector space and that we might even be able to identify cultural differences in aligned vectors with NVC.

The fastText vectors used in this work were sampled from two vector spaces, belonging to a set of over 40 aligned spaces. In the future, one should validate cross-lingual NVC for more languages and further explore how (and if) cultural differences can be identified with NVC.

Acknowledgements

This research was partially supported by the German Federal Ministry of Education and Research through the project DEEPLEE (01IW17001) and by Giance Technologies GmbH.

References

1. 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)
2. Brunet, M.E., Alkalay-Houlihan, C., Anderson, A., Zemel, R.: Understanding the origins of bias in word embeddings. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 803–811. PMLR, Long Beach, California, USA (09–15 Jun 2019)
3. Chen, K.J., Huang, C.R., Chang, L.P., Hsu, H.L.: Sinica corpus: Design methodology for balanced corpora. In: Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation. pp. 167–176 (1996)
4. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. arXiv preprint arXiv:1710.04087 (2017)
5. Dev, S., Phillips, J.: Attenuating bias in word vectors. In: Chaudhuri, K., Sugiyama, M. (eds.) Proceedings of Machine Learning Research. vol. 89, pp. 879–887. PMLR (16–18 Apr 2019)
6. Ghorbani, A., Wexler, J., Zou, J., Kim, B.: Towards Automatic Concept-based Explanations. Preprint at https://arxiv.org/abs/1902.03129 (2019)
7. Glavas, G., Litschko, R., Ruder, S., Vulic, I.: How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. pp. 710–721 (2019)
8. Gromann, D., Declerck, T.: Comparing pretrained multilingual word embeddings on an ontology alignment task. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). European Languages Resources Association (ELRA), Miyazaki, Japan (May 2018)

9. Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., Grave, E.: Loss in translation: Learning bilingual word mapping with a retrieval criterion. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2979–2984 (2018)
10. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: International Conference on Learning Representations (ICLR) (2015)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 (2013)
12. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168 (2013)
13. Prost, F., Thain, N., Bolukbasi, T.: Debiasing embeddings for reduced gender bias in text classification. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing. pp. 69–75 (2019)
14. Schwarzenberg, R., Raithel, L., Harbecke, D.: Neural vector conceptualization for word vector space interpretation. NAACL HLT 2019 (2019)
15. Wang, Z., Wang, H., Wen, J.R., Xiao, Y.: An Inference Approach to Basic Level of Categorization. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15. pp. 653–662. ACM Press (2015)