Predicting Concreteness and Imageability of Words Within and Across Languages via Word Embeddings

Nikola Ljubešić
Dept. of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
nikola.ljubesic@ijs.si

Darja Fišer
Dept. of Translation, Faculty of Arts
University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana
darja.fiser@ff.uni-lj.si

Anita Peti-Stantić
Faculty of Humanities and Social Sciences
University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb
anita.peti-stantic@ffzg.hr

Abstract

The notions of concreteness and imageability, traditionally important in psycholinguistics, are gaining significance in semantic-oriented natural language processing tasks. In this paper we investigate the predictability of these two concepts via supervised learning, using word embeddings as explanatory variables. We perform predictions both within and across languages by exploiting collections of cross-lingual embeddings aligned to a single vector space. We show that the notions of concreteness and imageability are highly predictable both within and across languages, with a moderate loss of up to 20% in correlation when predicting across languages. We further show that the cross-lingual transfer via word embeddings is more efficient than the simple transfer via bilingual dictionaries.

1 Introduction

Concreteness and imageability are very important notions in psycholinguistic research, building on the theory of the double, verbal and non-verbal, modality of representation of concrete words in the mental lexicon, contrasted to single verbal representation of abstract words (Paivio, 1975, 2010). Although often correlated with concreteness, imageability is not a redundant property. While most abstract things are hard to visualize, some call up images, e.g., *torture* calls up an emotional and even visual image. There are concrete things that are hard to visualize too, for example, *abbey* is harder to visualize than *banana* (Tsvetkov et al., 2014).

Both notions have proven to be useful in computational linguistics as well. Turney et al. (2011) present a supervised model that exploits concreteness to correctly classify 79% of adjective-noun pairs as having literal or non-literal meaning. Tsvetkov et al. (2014) exploit both the notions of concreteness and imageability to perform metaphor detection on subject-verb-object and adjective-noun relations, correctly classifying 82% and 86% instances, respectively.

The aim of this paper is to investigate the predictability of concreteness and imageability within a language, as well as across languages, by exploiting cross-lingual word embeddings as our available signal.

2 Related Work

While much work has been done on exploiting word embeddings in expanding sentiment lexicons (Tang et al., 2014; Amir et al., 2015; Hamilton et al., 2016), there is little work on predicting other lexical variables, concreteness and imageability included.

Tsvetkov et al. (2014) performed metaphor detection, using, among others, concreteness and imageability as their features. To propagate these features, obtained from the MRC psycholinguistic database (Wilson, 1988) to the entire lexicon, they used a supervised learning algorithm on vector space representations, where each vector element represented a feature. Performance of these classifiers was 0.94 for concreteness and 0.85 for imageability. They also applied the concreteness and imageability features to other languages by projecting features with bilingual dictionaries.

Broadwell et al. (2013) extended imageability scores to the whole lexicon by using the MRC

Proceedings of the 3rd Workshop on Representation Learning for NLP, pages 217–222
Melbourne, Australia, July 20, 2018.

imageability scores and hyponym and hyperonym links from WordNet.

Rothe et al. (2016) trained an orthogonal transformation to reorder word embedding dimensions into one-dimensional ultradense subspaces, the output thereby being a lexicon. They trained the transformations for sentiment, concreteness and frequency. For obtaining training data for concreteness, they used the BWK database (Brysbaert et al., 2014). They showed that concreteness and sentiment can be better extracted from embedding spaces than frequency, with a Kendall τ correlation coefficient of 0.623 for concreteness. Rothe and Schütze (2016) further exploited this method to perform operations over the extracted dimensions, such as given a concrete word like *friend*, find the related, but abstract word *friendship*.

Contributions In this paper we perform a systematic investigation of transfer of two lexical notions, concreteness and imageability, (1) to the remainder of the lexicon not covered in an annotation campaign, and (2) to other languages.

While there were already successful transfers within a language based on word embeddings (Tsvetkov et al., 2014; Rothe and Schütze, 2016), the only cross-lingual transfer was based on transfer via bilingual dictionaries (Tsvetkov et al., 2014). In this paper we compare the effectiveness of cross-lingual transfer via word embeddings and via bilingual dictionaries.

A byproduct of this research is a lexical resource in 77 languages containing per-word estimates for concreteness and imageability.

3 Data

3.1 Lexicons

In our experiments we use two existing English and one Croatian lexicon with concreteness and imageability ratings.

For English we use the MRC database (Wilson, 1988) (MRC onwards), consisting of 4,293 words with ratings for concreteness and imageability. The ratings range from 100 to 700 and were obtained by merging three different resources (Wilson, 1988).

We also use the BWK database consisting of 39,954 English words (Brysbaert et al., 2014) (BWK onwards) with concreteness ratings summarized through arithmetic mean and standard deviation. The ratings were collected in a crowdsourcing campaign in which each word was labeled by 20 annotators on a 1–5 scale.

For Croatian we use the MEGAHR database (MEGA onwards), consisting of 3,000 words, with concreteness and imageability ratings summarized through arithmetic mean and standard deviation. The ratings were collected in an annotation campaign among university students, with each word obtaining 30 annotations per variable on a 1–5 scale.

For performing cross-lingual transfer via a dictionary, we use data from a large popular online Croatian-English dictionary[1] containing around 100 thousand entries.

3.2 Embeddings

For both in-language and cross-lingual experiments we use the aligned Facebook collection of embeddings[2], trained with fastText (Bojanowski et al., 2016) on Wikipedia dumps, with embedding spaces aligned between languages with a linear transformation learned via SVD (Smith et al., 2017) on a bilingual dictionary of 500 out of the 1000 most frequent English words, obtained via the Google Translate API[3].

We also experimented with another cross-lingual embedding collection (Conneau et al., 2017), obtaining similar results and backing all our conclusions. This is in line with recent work on comparing cross-lingual embedding models which suggests that the actual choice of monolingual and bilingual signal is more important for the final model performance than the actual underlying architecture (Levy et al., 2017; Ruder et al., 2017). Given that one of our goals is to transfer concreteness and imageability annotations to as many languages as possible, using cross-lingual word embeddings based on Wikipedia dumps and dictionaries obtained through a translation API is the most plausible option.

4 Experiments

4.1 Setup

We perform two sets of experiments: one within each language, and another across languages.

[1] http://www.taktikanova.hr/eh/

[2] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

[3] https://github.com/Babylonpartners/fastText_multilingual

While in-language experiments are always based on supervised learning, in cross-lingual experiments we compare two transfer approaches: one based on a simple dictionary transfer, and another on supervised learning on the word embeddings in the source language, and performing predictions on word embeddings in the target language, with the two embedding spaces being aligned.

We perform our prediction experiments by training SVM regression models (SVR) and deep feedforward neural networks (FFN) over standardized (zero mean, unit variance) embeddings and each specific response variable. We experiment with all available gold annotations as our response variables, namely both the arithmetic mean and standard deviation of concreteness and imageability.

We tuned the hyperparameters of each of the regressors on a subset of the Croatian, MEGA dataset in the case of the in-language experiments, and another subset of the BWK dataset for the cross-lingual experiments. Given that we perform the final experiments on the whole datasets, and that we have two additional English datasets at our disposal for the in-language experiments and three additional dataset pairs for the cross-lingual experiments, we consider our approach to be resistant to the overfitting of the hyperparameters going unnoticed.

While the SVR proved to work well with the RBF kernel, the C hyperparameter of 1.0 and the γ hyperparameter of 0.003, the feedforward network obtained strong results with two fully-connected hidden layers, consisting of 128 and 32 units each and ReLU activation functions, with a dropout layer after each of the hidden layers, and an output layer with a linear activation function. We optimized for the mean squared error loss function and ran 50 epochs on each of the datasets, with a batch size of 32.

While we used the same regressor setup for the SVR system for both the in-language and cross-lingual experiments, for the FFN system the dropout probability in the in-language experiments was 0.5, while in the cross-lingual setting the dropout probability was set to 0.8, obtaining thereby a more general model which transfers better to the other language.

We perform in-language experiments via 3-fold cross-validation, while we train models on our source language dataset and evaluate the models on our target language dataset for cross-lingual experiments. We evaluate each approach via the Spearman rank and Pearson linear correlation coefficients. In the paper we report the Spearman correlation coefficient only as the relationships across both metrics in all the experiments are identical. We perform our experiments with the `scikit-learn` (Pedregosa et al., 2011) and `keras` (Chollet et al., 2015) toolkits.

4.2 In-language Experiments

We start our experiments in the in-language setting, running cross-validation experiments over each of our three datasets on all available variables. The results of these experiments, with some basic information on the size of the datasets, are given in Table 1. Aside from the three lexicons introduced in Section 3.1, we experiment with another lexicon, BWK.3K, which is a randomly downsampled version of the BWK lexicon to the size of the two remaining lexicons. We introduce this additional resource (1) to control for dataset size when comparing results on our different datasets and (2) to measure the impact of training data size by comparing the results on the two flavours of the BWK dataset.

The results in Table 1 show that the support vector regressor consistently performs better than the feedforward neural network at predicting almost all values, with relative error reduction lying between 7% and 12%. The bold results are statistically significantly better than the corresponding non-bold ones given the approximate randomization test (Edgington, 1969) with $p < 0.05$. Our assumption is that the stronger FFN model does not show a positive impact primarily due to the small size of the datasets and the simplicity of the modeling problem.

We can further observe that the arithmetic mean is much easier to predict than standard deviation on both variables in all the datasets. This can be explained by the fact that standard deviation on the two phenomena can partially be explained with the level of ambiguity of a specific word, and this type of information is at least not directly available in context-based word embeddings.

Furthermore, imageability seems to be consistently slightly harder to predict than concreteness. Our initial assumption regarding this difference was that imageability is a more vague notion for

dataset	MEGA		BWK		BWK.3K		MRC	
lang	hr		en		en		en	
size	2,682		22,797		3,000		4,061	
method	SVR	FFN	SVR	FFN	SVR	FFN	SVR	FFN
C.M	**0.760**	0.742	**0.887**	0.879	**0.848**	0.834	**0.872**	0.863
C.STD	**0.265**	**0.274**	**0.484**	0.461	**0.376**	**0.364**	-	-
I.M	**0.645**	0.602	-	-	-	-	**0.803**	0.787
I.STD	**0.439**	**0.415**	-	-	-	-	-	-

Table 1: Results of the in-language experiments on predicting mean (.M) and standard deviation (.STD) of concreteness (C) and imageability (I), either using a support vector regressor (SVR) or feed-forward network (FFN). Evaluation metric is the Spearman correlation coefficient.

human subjects, and therefore their responses are more dispersed, adding to the complexity of the prediction. However, analyzing standard deviations over concreteness and imageability showed that these are rather the same. We leave this open question for future research.

When comparing the results on predicting mean concreteness on the full BWK and the trimmed BWK.3K datasets, we see a significant improvement of the predictions of the on the larger dataset, showing that having 10 times more data for learning can produce significant improvements in the prediction quality.

4.3 Cross-lingual Experiments

In cross-lingual experiments we compare our two approaches to cross-lingual transfer: dictionary lookup (DIC onwards) and supervised learning on aligned word embedding spaces via the two methods introduced in Section 4.2, SVR and FFN.

The DIC method simply looks up for each word in the source language resource all possible translations to the target language and directly transfers the concreteness and imageability ratings to the target language words. In case of collisions in the target language (two source language words being translated to the same word in the target language), we perform averaging over the transfered ratings. In our experiments, the arithmetic mean showed to be a better averaging method than the median, we therefore report the results on that averaging method.

The SVR and FFN methods use supervised learning in a very similar fashion to the in-language experiments described in Section 4.2. We train a supervised regression model on the whole source language dataset, using word embedding dimensions as features and the variable of choice as our target. We obtain estimates of our variable of choice in the target language by applying the source-language model on the target-language word embeddings since the two embedding spaces are aligned.

For both approaches we compare the target-language estimates with the gold data available from our lexicons.

We present the results of the cross-lingual experiments in Table 2. Our first observation is that, while in the in-language setting the SVR method has regularly outperformed the FFN method, in the cross-lingual setting this is not the case any more, with SVR and FFN obtaining very similar results, in five out of six cases in the range of no statistically significant difference. Our explanation for the loss of the positive impact in using the weaker, support vector regression model, is that with the noisy alignment of the two embedding spaces the prediction problem became harder, now both models performing similarly. While the strong point of SVR is that it performs very well on small datasets, the strong point of the FFN method is that it generalizes better.

That higher generalization is beneficial in case of the cross-lingual problem is observable in the difference in the hyperparameter tuning results on the FFN method, where in the in-language setting the optimal dropout was 0.5, while in the cross-lingual setting it is 0.8.

Our second observation is that all the predicted ratings suffer in the cross-lingual setting, when compared to the in-language results presented in Table 1, observing for the SVR method a drop of around 5 to 15%. While standard deviation was already poorly predicted in the in-language set-

source target	MEGA (hr) BWK (en)			BWK (en) MEGA (hr)			MEGA (hr) MRC (en)			MRC (en) MEGA (hr)		
	SVR	FFN	DIC	SVR	FFN	DIC	SVR	FFN	DIC	SVR	FFN	DIC
C.M	**0.791**	**0.793**	0.728	**0.724**	**0.719**	0.641	**0.797**	**0.794**	0.611	**0.651**	**0.644**	0.638
C.STD	0.178	0.141	0.224	0.185	0.145	0.137	-	-	-	-		
I.M	-	-	-	-	-	-	**0.694**	**0.683**	0.523	**0.548**	0.531	0.503

Table 2: Results of the cross-lingual experiments, either using supervised learning (SVR, FFN), or simple dictionary lookup (DIC). Evaluation metric is the Spearman correlation coefficient. Results in bold are best results per problem with no statistically significant difference.

ting, in the cross-lingual setting it drops even further to a non-useful level, below 0.2. This is the reason why we do not calculate statistical significance of the differences in these results and do not include their estimates in our final 77-languages-strong resource. In the final cross-lingual resource we include only the mean of concreteness and imageability, the notions for which we have obtained strong correlation in our cross-lingual experiments.

Finally, when comparing the cross-lingual transfer via embeddings (SVR and FFN) and via a dictionary (DIC), the learning-on-embeddings approach outperforms the dictionary method in each instance, with the relative loss in correlation when moving from the EMB to the DIC approach of 5% to 25%.

4.4 Regressor Coefficient Analysis

Our final analysis concerns the question of how many of the embedding dimensions are crucial for our regressors to predict the notions of concreteness and imageability. We consider two potential scenarios: (1) each of the notions are encoded in one or a few of the embedding dimensions and (2) the notions are encoded in many embedding dimensions.

The analysis is performed by calculating the cumulative distribution of absolute and normalized (sum to 1), reversely sorted coefficients of the SVM regressor with a linear kernel. For both phenomena, concreteness and imageability, the distributions show that the predictions are based on a significant number of embedding dimensions. Namely, while 80 most informative dimensions cover 50% of the coefficients' mass, half of the dimensions (150) cover 80% of that mass. This shows for the second scenario – concreteness and imageability are encoded in a significant number of embedding dimensions – to be true.

5 Conclusion

In this paper we have shown that concreteness and imageability ratings can be successfully transfered both to non-covered portions of the lexicon and to other languages via (cross-lingual) word embeddings.

With the in-language experiments we have shown that the arithmetic mean of both notions is much easier to predict than their standard deviation, the latter probably encoding word ambiguity, type of information not directly present in word embeddings.

Our experiments across languages have shown that the loss in comparison to in-language experiments on predicting the means of both concreteness and imageability are around 15%, a reasonable price to pay given the applicability of the method to all of the 77 languages present in the word embedding collection. The predictions of concreteness and imageabililty obtained in the 77 languages are available at http://hdl.handle.net/11356/1187.[4]

Comparing the two methods of transfer – dictionary vs. cross-lingual embeddings, shows regularly better (5%–15%) results of the latter, proving once more the usefulness of word embeddings, especially in the currently expanding cross-lingual setup.

Acknowledgements

The work described in this paper has been funded by the Croatian National Foundation project HRZZ-IP-2016-06-1210, the Slovenian Research Agency project ARRS J7-8280, and by the Slovenian research infrastructure CLARIN.SI.

[4]Ongoing developments are stored at https://github.com/clarinsi/megahr-crossling/.

References

Silvio Amir, Ramón Astudillo, Wang Ling, Bruno Martins, Mario J Silva, and Isabel Trancoso. 2015. Inesc-id: A regression model for large scale twitter sentiment lexicon induction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. pages 613–618.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

George Aaron Broadwell, Umit Boz, Ignacio Cases, Tomek Strzalkowski, Laurie Feldman, Sarah Taylor, Samira Shaikh, Ting Liu, Kit Cho, and Nick Webb. 2013. Using imageability and topic chaining to locate metaphors in linguistic corpora. In Ariel M. Greenberg, William G. Kennedy, and Nathan D. Bos, editors, *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer Berlin Heidelberg, Berlin, Heidelberg, pages 102–110.

Marc Brysbaert, AB Warriner, and V Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *BEHAVIOR RESEARCH METHODS* 46(3):904–911.

François Chollet et al. 2015. Keras. `https://keras.io`.

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

Eugene S. Edgington. 1969. Approximate randomization tests. *The Journal of Psychology* 72(2):143–149. https://doi.org/10.1080/00223980.1969.10543491.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. pages 595–605. http://aclweb.org/anthology/D/D16/D16-1057.pdf.

Omer Levy, Anders Søgaard, and Yoav Goldberg. 2017. A strong baseline for learning cross-lingual word embeddings from sentence alignments. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, pages 765–774. http://aclweb.org/anthology/E17-1072.

A. Paivio. 1975. *Coding Distinctions and Repetition Effects in Memory*. Research bulletin. Department of Psychology, University of Western Ontario.

Allan Paivio. 2010. Dual coding theory and the mental lexicon. *The Mental Lexicon* 5(2):205–230.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. *CoRR* abs/1602.07572.

Sascha Rothe and Hinrich Schütze. 2016. Word embedding calculus in meaningful ultradense subspaces. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. volume 2, pages 512–517.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A survey of cross-lingual embedding models. *CoRR* abs/1706.04902. http://arxiv.org/abs/1706.04902.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR* abs/1702.03859. http://arxiv.org/abs/1702.03859.

Duyu Tang, Furu Wei, Bing Qin, Ming Zhou, and Ting Liu. 2014. Building large-scale twitter-specific sentiment lexicon: A representation learning approach. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. pages 172–182.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *ACL*.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '11, pages 680–690.

Michael Wilson. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers* 20(1):6–10.