Identifying the Importance of Content Overlap for Better Cross-lingual Embedding Mappings

Réka Cserháti and **Gábor Berend**
Department of Computer Science
University of Szeged
{cserhatir,berendg}@inf.u-szeged.hu

Abstract

In this work, we analyze the performance and properties of cross-lingual word embedding models created by mapping-based alignment methods. We use several measures of corpus and embedding similarity to predict BLI scores of cross-lingual embedding mappings over three types of corpora, three embedding methods and 55 language pairs. Our experimental results corroborate that instead of mere size, the amount of common content in the training corpora is essential. This phenomenon manifests in that i) despite of the smaller corpus sizes, using only the comparable parts of Wikipedia for training the monolingual embedding spaces to be mapped is often more efficient than relying on all the contents of Wikipedia, ii) the smaller, in return less diversified Spanish Wikipedia works almost always much better as a training corpus for bilingual mappings than the ubiquitously used English Wikipedia.

1 Introduction

Word embedding methods (e.g. Mikolov et al., 2013b, Pennington et al., 2014, Bojanowski et al., 2017) have become an essential tool for representing words in most NLP tasks. These algorithms assign a low-dimensional vector to words based on the patterns of their contexts in a training corpus, and this way they locate the words in the vector space in a consistent way, so that words with similar meaning are assigned to similar vectors. Therefore, it can be assumed that the layout of the word vectors are near equivalent in independently trained models, so word embedding models in different languages can be aligned into a common space. Such alignments are a standard way of creating bi- or multilingual word embedding spaces, which are very useful for machine translation and a wide range of cross-lingual NLP tasks.

Although large pre-trained language models are superior to traditional word embeddings in many NLP tasks, one strength of these mapping-based methods is their extensive applicability, e.g. for low-resource languages or special domain (e.g. medical) data. Additionally, probably due to significantly lower resource requirements (Strubell et al., 2019) and often competitive results (Litschko et al., 2021), a large proportion of industrial NLP applications is still based on static word embeddings (Arora et al., 2020).

On the other hand, the results of the mappings and the performance of the multilingual models was still shown to be extremely dependent on the mapping scenario (Søgaard et al., 2018; Vulić et al., 2019, 2020). Previous work attributes low performance and non-isomorphism of monolingual embedding models to typological differences between languages, domain differences in the training corpora, insufficient resources, and under-training (Doval et al., 2020; Søgaard et al., 2018; Vulić et al., 2020).

We rely on popular bilingual alignment methods, and conduct a thorough analysis of the connection between the evaluation scores of the mappings and language relatedness, isomorphism of the source embeddings measured by several metrics, and some newly identified, easy to calculate corpus properties that are highly predictive of the bilingual mapping performance: the token overlap ratio in the vocabularies, and the distance between word distributions of the corpora. These combined with corpus size surpass existing isomorphism measures as predictors of bilingual mapping score.

Our goal is to help researchers and developers use resources more efficiently, and find the most appropriate setting for creating bilingual word embedding models. [1]

[1] Our codes, mapping dictionaries, and more mapping results are available at https://github.com/xerevity/mappability

2 Related Work

2.1 Mapping Algorithms

The success of the pioneering neural word embedding models (Mikolov et al., 2013b) almost immediately led to the idea of creating bilingual models using linear transformations (Mikolov et al., 2013a). The original problem is finding a mapping that transforms an embedding matrix close to another in a different language. Mikolov et al. (2013a) solve this with stochastic gradient descent, minimizing the squared euclidean distance of word pairs from a seed dictionary.

Subsequently, several works improved this mapping method: for example, Xing et al. (2015) normalize the source and target embeddings, and constrain the mapping to be orthogonal; Artetxe et al. (2016) center the mean, and find the transformation in closed form, solving a least squares problem. Later Artetxe et al. (2018a) proposed a multi-step framework consisting of mean-centering, normalization, whitening, and an orthogonal transformation. In contrast, *RCSLS* (Joulin et al., 2018) is based on relaxing the orthogonal restriction and returning to stochastic gradient descent with a different loss function, aiming to be consistent with the *CSLS* (Cross-domain Similarity Local Scaling; Conneau et al., 2018) retrieval method.

Additionally, aligning word embeddings in an unsupervised way, without any cross-lingual signal has also become an exciting topic, giving rise to diverse approaches of unsupervised embedding mappings. The first really successful solution by Conneau et al. (2018) is based on adversarial learning; Artetxe et al. (2018b) proposed an iterative self-learning method initialized by sorting embedding values; and Non-adversarial Translation by Hoshen and Wolf (2018) also uses self-learning, but a different method, initialized using PCA.

2.2 Analysis of Cross-Lingual Word Embeddings

Several works have already analyzed performance of cross-lingual embedding mappings (e.g. Kementchedjhieva et al., 2019; Glavaš et al., 2019; Vulić et al., 2019). More related to this paper, reasons why some settings do not work well were also investigated. For instance, Søgaard et al. (2018) use eigenvector similarity of nearest neighbor graphs to show that the isomorphic assumption does not hold in many cases, and report the negative effect of language and domain dissimilarity on the unsupervised embedding alignment method of Conneau et al. (2018).

In addition, Vulić et al. (2020) states that small corpora and under-training also play a significant role in non-isomorphism of word embeddings. Dubossarsky et al. (2020) also examine isomorphism, and suggest some new measures to quantify transferability of embedding spaces based on their spectral statistics: how similar their singular values are on the one hand, and their individual robustness measured by condition numbers, on the other hand.

In the rest of this paper, we supplement isomorphism measures with corpus similarity measures, and show that corpus similarity is one of the key factors influencing mapping scores in both supervised and unsupervised cases. We show that two corpora of sufficient size, coming from the same domain (Wikipedia in this case) can still be too different for good mapping scores, while good mappings are possible on relatively small corpora if other important conditions are met.

3 Experimental Setup

3.1 Corpora

In our experiments, we compare BLI (Bilingual Lexicon Induction) scores of embeddings trained on three types of corpora, all of them extracted from Wikipedia:

1. We use the full Wikipedia texts[2] of our 11 languages studied: Czech, Danish, German, Greek, English, Spanish, Finnish, Hungarian, Norwegian, Romanian and Turkish. These all come from the same domain (encyclopedia articles), which condition was reported to be necessary for sufficient unsupervised mappings by Søgaard et al. (2018). Nevertheless, this type is the least restricted in our experiments, and the sizes may be very dissimilar for different languages, but these are the largest among our experiments. We call this type of corpora loosely-comparable Wikipedia, or **L-Wiki** for short.

2. Even within the same domain, the content of the corpora may be very different, which might (and, according to our hypothesis, does) have a negative influence on the mappings. Therefore, we create a mildly-comparable (**M-Wiki**) corpus, separately for all of our 55 lan-

[2] As available at: https://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/

guage pairs, by filtering articles with bidirectional cross-language links between the two Wikipedias. This is also expected to make sizes comparable within a language pair, but the different length of the articles may still cause dissimilarity in size. Additionally, the amount of filtered parts between different language pairs is especially variable.

3. In terms of both size and content, a parallel corpus between two languages is as similar as possible. As such, we use the WikiMatrix (Schwenk et al., 2021) parallel corpus (hereafter strictly-comparable Wiki or **S-Wiki**), which is also extracted from Wikipedia. The sizes here are substantially smaller than in the previous types, but also vary by language pairs.

This way, we have various levels of corpus size, language relatedness, and proportion of overlapping information among our experimental language pairs. To the best of our knowledge, our experiments are the first to analyze different corpus types, and to dissect the effects of corpus similarity on the quality of bilingual embedding mapping.

3.2 Training and Test Dictionaries

Since there is no available gold standard dictionary for most of our language pairs, we create silver dictionaries from the WikiMatrix parallel corpora using the word2word (Choe et al., 2020) tool, which generates translations for words based on parallel sentences. To ensure the quality of these, we generate two translations for each word above the mean frequency in the corpus, and only keep pairs that are mutual translations of each other. Then we randomly select (disjoint) training and test dictionaries with 3000 and 1000 source words, respectively.

3.3 Word Embedding Models

We train FastText (Bojanowski et al., 2017) word vectors on all of the above corpora using Gensim (Řehůřek and Sojka, 2010), with the following hyperparameters: dimensions: 300, negative samples: 5, context window: 5, minimum word count: 5, maximum vocabulary size: 200 000.

To create the cross-lingual models we use three mapping methods: supervised VecMap (Artetxe et al., 2018a), RCSLS (Joulin et al., 2018), and Non-adversarial Translation (NAT; Hoshen and Wolf, 2018) on the embeddings trained on three

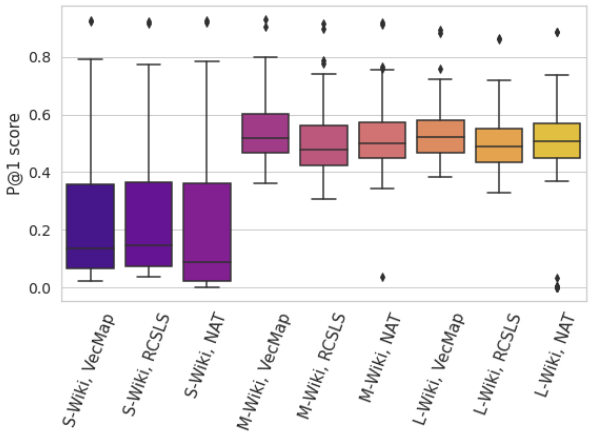


Figure 1: Distributions of BLI scores of embedding mappings using different methods and corpus types.

types of corpora and 110 language pairs, performing a total of 990 mappings as we separate different source–target directions of the same language pair.

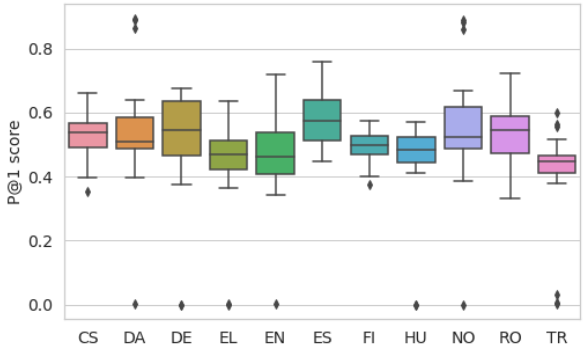
We evaluate the models with P@1 scores, i.e. by finding the nearest neighbor of a source word among the target language embeddings, and see whether it is a correct translation according to our dictionary. We experimented with other, more sophisticated evaluation methods as well, but the scores did not change relative to each other.

4 Results and Analysis

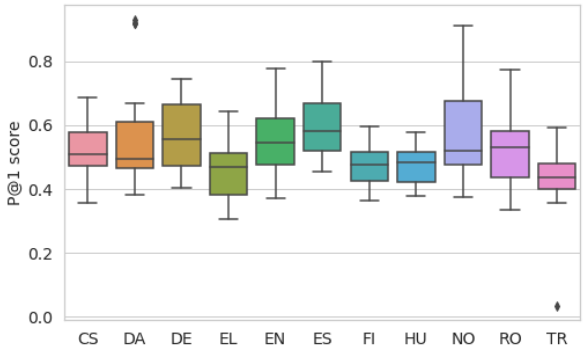
4.1 Mapping Methods

We show the distributions of the used mapping algorithms on separate corpus types in Figure 1. Our first important observation is that the results are much more dependent on the corpus type than on the mapping algorithm. While mappings of mildly and loosely comparable corpora reach similar median scores and extremes, the strictly-comparable (hence a lot smaller) corpus mapping scores range much wider. The median of the S-Wiki scores is very low, but the highest scores and quantiles are in line with the other corpus types. Later we will also investigate in which cases do these mappings perform well, and why.

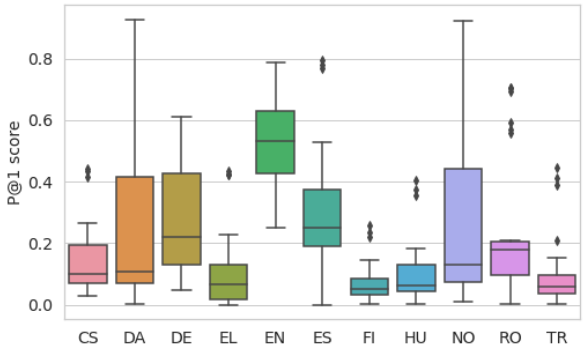
Another information visible in Figure 1 is that in our settings, VecMap performs the best among these three algorithms. However, except some cases where it completely fails, Non-adversarial Translation reaches competitive results to the other methods, despite the lack of supervision.



(a) Score distributions of languages involved in mapping embeddings trained on loosely-comparable full Wikipedia.



(b) Mapping scores of languages, with embeddings trained on mildly-comparable Wiki corpus.



(c) Mapping scores of languages, with embeddings trained on a strictly-comparable (parallel) corpus.

Figure 2: Score distributions of embedding mappings involving a language (either as source or as target language).

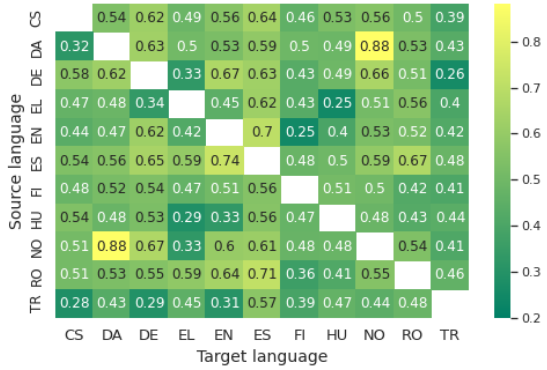


Figure 3: Average P@1 scores of loosely-comparable Wikipedia embedding mappings in all examined source–target pairs of languages.

4.2 Languages

The effect of the languages used on the performance of the cross-lingual embeddings has been widely studied (Vulić et al., 2019; Dubossarsky et al., 2020; Doval et al., 2020), but our evaluation on 110 pairs of 11 languages still shows interesting and instructive patterns. It is conspicuous in Figure 2a that, despite the widespread use of English as a transfer language, mappings of loosely-comparable Wikipedia embeddings involving Spanish perform substantially better. In this case, English Wikipedia probably covers too diverse and deep articles that none of the other Wikipedias do, which makes Spanish Wikipedia a better corpus for embedding mappings. However, using mildly-comparable Wikipedia weakens this phenomenon (see Figure 2b), which might suggest that instead of the corpus size, the real indicator of performance is the amount of overlapping information between the two corpora. We will deal with this hypothesis a lot more in the rest of this article.

Figure 3 shows the average performance of loosely-comparable Wikipedia mappings broken down by both source and target language. Beside some other interesting details, an outstanding result of the Danish–Norwegian mapping is clear. In this case, language relatedness, geographical and cultural similarity are all given, therefore we can assume that the two Wikipedias are also very similar in topics, style, editing, etc. This can be considered a case where all the necessary factors are met for obtaining a high-performing mapping.

From this figure it seems that only very close language relatedness is really beneficial, e.g. between Germanic and Romance languages, but Germanic

	L-Wiki	M-Wiki	S-Wiki
Token overlap	0.458	**0.747**	**0.942**
Word distributions	**–0.603**	–0.708	–0.616
Size (log)	0.306	0.552	0.877

Table 1: Pearson correlation coefficients between P@1 scores and corpus properties.

languages, for example, can be mapped to linguistically very distant Finno-Ugrian languages just as well as to non-Germanic Indo-European languages, which might also be a useful observation for future work.

4.3 Corpus Size and Similarity

Our key observation is that one of the most required condition for good embedding mappings is corpus similarity, more precisely the amount of common contexts the words appear in, as a complement to previous claims pointing to language similarity and corpus size (Dubossarsky et al., 2020; Vulić et al., 2020).

We introduce two measures to quantify corpus similarity:

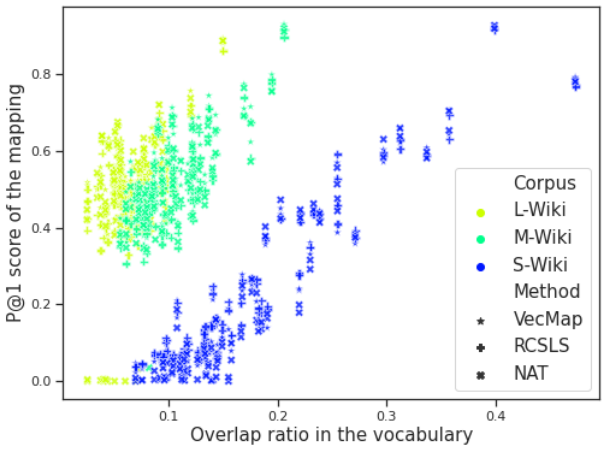
- *Token Overlap* is the ratio of token forms used in both corpora to the number of tokens used in one or both of them.

$$TO(V_1, V_2) = \frac{|V_1 \cap V_2|}{|V_1 \cup V_2|}$$

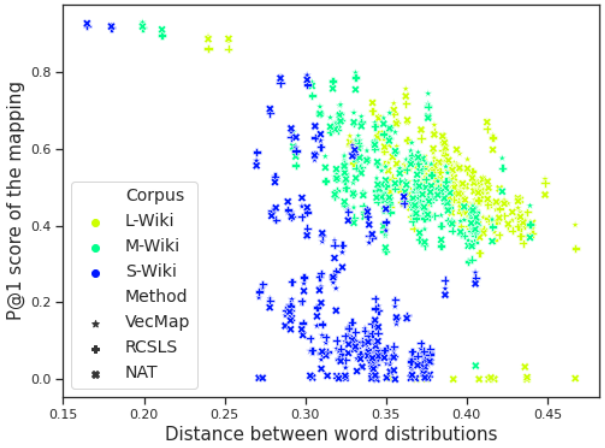
  Most of these shared tokens are probably words of foreign origin, having the same meaning, therefore their presence in large proportions indicates similar content in the texts. However, this measure is affected by language similarity as well, and is unusable with languages written in different scripts.

- As *Word Distribution Distance* between two corpora we take the normalized frequency distribution of words from our silver dictionary between the languages of the two corpora, and compute Jensen–Shannon divergence between them. This way, we use the words of the dictionary as keywords, and the divergence will be small only if the respective topics appear in a similar proportion in the corpora.

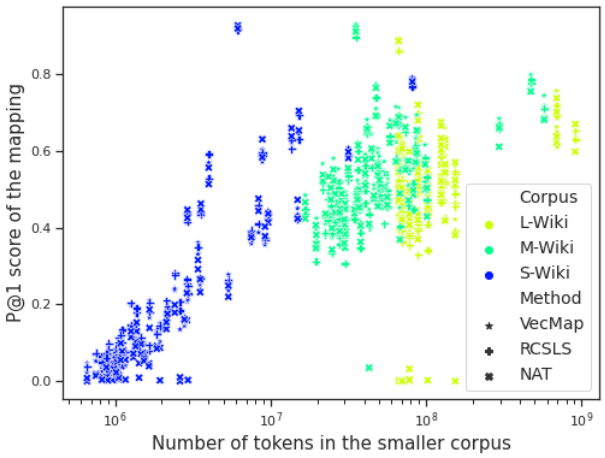
As Vulić et al. (2020) showed, mapping scores are greatly influenced by the size of the training corpora, therefore we include this information to



(a) Connection of mapping score to the proportion of overlapping tokens



(b) Connection of mapping score to word distribution distance



(c) Connection of mapping score to corpus size (on a logarithmic scale)

Figure 4: Relationship between performance of bilingual mappings and corpus properties.

our corpus data as well. We examined correlations using the token numbers of the source and the target corpus, the arithmetic and harmonic mean of them, and the minimum of them. The latter of these proved to be the most powerful indicator of mapping scores, so we use the minimum of token numbers of the two training corpora involved in the mapping to represent corpus size.

Figure 4 shows the connection of the mapping scores to the above defined corpus similarity measures and corpus size. All of these corpus properties seem to indicate performance well, as models with more overlap in their vocabularies, with more similar word distributions, or trained on a larger corpora perform generally better.

However, the parameters of the regression lines for corpus types differ clearly, implying that when we make the corpora mildly and strictly comparable, so overlap ratio and word distribution similarity increase, the results are not improving as much as we could have expected by extrapolating the scores of the full Wikipedia corpus. But similarly, the smaller size of comparable and parallel corpora does not directly lead to a decrease in performance either.

It can be clearly seen from Figure 4c, that although there is a connection between corpus size and cross-lingual mapping score as well, big corpora are neither necessary nor sufficient for good results: some of the best scores are reached by mapping embeddings trained on the Danish–Norwegian parallel corpus, having less than 10 million tokens, while the biggest corpora consist of approximately 1 billion tokens.

Also, correlations in Table 1 show that the relationship between corpus size and performance gets stronger as the corpus is filtered for overlapping articles (M-Wiki), and even stronger for parallel sentences (S-Wiki), again supporting our hypothesis that the amount of common information is an important factor for mappings. We measured corpus similarity by the number of common tokens and the distribution of dictionary words successfully, but there are probably other, more widely usable measures to be found in future work.

4.4 Content Overlap

To further validate our statement that common content in corpora greatly influences mapping scores, we conduct controlled experiments, in which we align embeddings of the same language, and construct the training corpora from subsets of a single Wikipedia. For 3 languages (English, Spanish and Hungarian) we create corpora in different sizes, and for all of these we select subsets that are matching in size, but contain 0%, 33%, 66%, or 100% of the text in the first corpus.

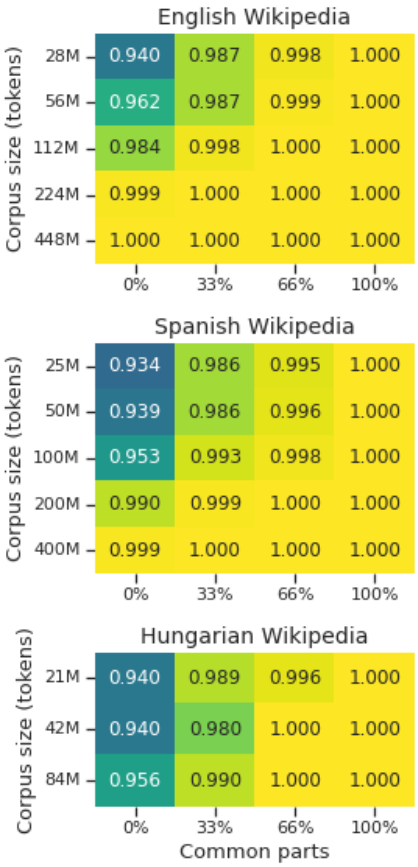


Figure 5: P@1 scores of RCSLS mappings of embeddings from Wikipedia parts of various overlap ratio and size.

This methodology allows us to examine the effects of size and content overlap explicitly (and exclude the effects of typological differences between languages). These parts, however, may contain very similar articles in the same field, which are not accounted as overlap. Probably this is why small size and zero overlap still yield very high P@1 scores, as shown in Figure 5. Still, the trends are convincing that content overlap is at least as important as corpus size. We show the scores of the RCSLS mapping, but the same patterns can be seen with other methods.

These results also imply that word embeddings represent word usage of a specific corpus, rather than a whole language, which is often forgotten in multilingual tasks. Therefore, it is possible that a

corpus can even be too large compared to another if there are too many different contexts appearing in only one of them, which might explain why Spanish Wikipedia is superior to English as a training corpus. We can conclude that even among corpora of the same domain, corpus similarity is a major requirement for the success of word embedding alignment.

4.5 Embedding Isomorphism

Previous work has extensively studied the (non-) isomorphism of word embeddings, and its effect on bilingual alignments. This problem can be considered one of the core questions of bilingual mappings, since this method gains its inspiration and theoretical validity from the assumption that embeddings trained on different languages should be approximately isomorphic. However, our surprising results show that the degree of isomorphism is generally less correlated to BLI scores than corpus properties.

Measuring the degree of isomorphism between word embedding models is an interesting question in itself as well, and several solutions have already been proposed for it. We adopt five existing measures (for more details see the works cited below) and introduce a new one, based on the similarity of words.

	L-Wiki	M-Wiki	S-Wiki
Laplacian	–0.557	**–0.777**	**–0.396**
SVG	–0.174	–0.275	0.157
Spectral	–0.392	–0.290	0.061
Effective Spectral	–0.251	–0.236	–0.224
Relational	**0.591**	0.486	0.262
Neighbors	0.521	0.501	–0.099

Table 2: Pearson correlation coefficients between P@1 scores and isomorphism.

	L-Wiki	M-Wiki	S-Wiki
#	62	72	27
Laplacian	–0.851	–0.681	**–0.926**
SVG	–0.346	–0.518	–0.896
Spectral	–0.383	0.170	–0.548
Effective Spectral	–0.174	0.118	0.265
Relational	**0.895**	**0.879**	0.921
Neighbors	0.839	0.704	0.898

Table 3: Pearson correlation coefficients between P@1 scores and isomorphism among mappings scoring 0.6 or higher. We indicate the number of mappings meeting this criterion below the corpus types.

- *Laplacian Isospectrality* (Søgaard et al., 2018) measures the difference between the Laplacian eigenvalues of word nearest neighbor graphs. We take the average isospectrality of 10 graphs, each constructed of 50 random words from our dictionary and their translations.

- *Singular Value Gap* (SVG; Dubossarsky et al., 2020) is the distance between the sorted singular values of the two word embedding matrices.

- *Spectral Condition* (Dubossarsky et al., 2020) is the harmonic mean of the condition numbers of the two embedding matrices, which measure their sensitivity to noise.

- *Effective Spectral Condition* (Dubossarsky et al., 2020) is the harmonic mean of the effective condition numbers of the two embedding matrices.

- *Relational Similarity* (Vulić et al., 2020) quantifies how similarly the two models rate the proximity of word pairs. We take 10,000 random word pairs and their translations from our dictionary, and compute the Pearson correlation coefficient between the two lists of cosine similarity scores between the pairs.

- *Neighbor Overlap* quantifies the overlap between the neighborhood of words in the embedding models. We take a word from the dictionary in one language, and find its 10 nearest neighbors among the dictionary entries. Then we count how many of the translations of these appear among the nearest neighbors of the translation of the original word. We repeat this 1000 times, and compute the average of the outcomes.

In Table 2 we show the correlations between the above isomorphism measures and mapping scores. It is interesting that while the connection between corpus properties and bilingual scores was the strongest in strictly-comparable corpora, the opposite seems to be true in this case: the performance of mapping embeddings trained on S-Wiki seems not to be very dependent on isomorphism. Often it even happens that the correlation to embedding similarity/dissimilarity turns into negative/positive in the strictly-comparable case.

This raises the question if it is possible that two models, in which the same word has different neighbors, are transformed so that the appropriate words still become nearest neighbors, or above a certain score isomorphism remains a requirement for bilingual performance. To answer this, we compute the correlations between isomorphism measures and mapping scores again, but only among mappings with P@1 score 0.6 or higher. Table 3 shows that in these cases performance does indeed depend on isomorphism, especially on Laplacian, relational, and neighbor similarities.

We can see that mapping scores are connected to the isomorphism of source embeddings, especially among the relatively well performing models. Therefore we can use both isomorphism and corpus similarity to predict bilingual performance, which we will do in the next section.

4.6 Predicting BLI Scores

In our final experiments, we try to predict the mapping scores from the above studied corpus and isomorphism measures. We make predictions based on our three corpus properties, six isomorphism measures, and all of these combined, using random forest regression with the default parameters in Scikit-learn (Pedregosa et al., 2011), evaluating the model with the Leave-One-Out method.

The results in Table 4 show that mapping scores are very well predictable in most cases, but this varies between corpus types and alignment methods. Properties of the corpus, however, are almost always better predictors than isomorphism; the only exception is Non-adversarial Translation of loosely-comparable Wikipedia embeddings.

Combining corpus and isomorphism measures usually does not lead to an improvement either, which could mean that isomorphism depends on corpus properties as well. To find this out, we make predictions of all isomorphism measures from our three corpus properties, and show the results in Table 5. From these we see that although isomorphism does not depend solely on corpus similarity, it is also greatly influenced by it.

It is important to note that we did not use language information at all, therefore these high scores mean that the above corpus measures are more important than language similarity, or at least they carry this information as well. These results again support our observation on the importance of corpus similarity for good performance of bilingual word embedding mappings.

L-Wiki:			
	VecMap	**RCSLS**	**NAT**
Corpus	0.770	0.709	0.366
Isomorphism	0.720	0.654	0.481
All	0.788	0.733	0.478
M-Wiki:			
	VecMap	**RCSLS**	**NAT**
Corpus	0.797	0.794	0.664
Isomorphism	0.672	0.670	0.581
All	0.799	0.780	0.658
S-Wiki:			
	VecMap	**RCSLS**	**NAT**
Corpus	0.975	0.981	0.971
Isomorphism	0.777	0.793	0.765
All	0.972	0.974	0.966

Table 4: R2 scores of the predictions of P@1 scores, using random forest regression based on our three corpus properties combined, six isomorphism measures combined, and all of these.

5 Conclusion

We examined the connection of embedding mapping scores to languages, corpus properties, and embedding isomorphism. We found that the Spanish Wikipedia is better for this purpose than the English Wikipedia, often used by default. This is explained by our other experiments on the relationship of corpus properties and mapping quality, where it turned out that corpus similarity is at least as important as corpus size, therefore the hugeness and wide diversity of the English Wikipedia can be harmful.

Moreover, we have seen that language similarity is really beneficial for very closely related languages only, e.g. between Germanic or Romance languages. Mapping scores are well predictable even without any information about the languages, based on three properties of the corpora: corpus size, proportion of common tokens, and distance of the word distributions. These data also surpass existing embedding isomorphism measures as predictors.

On the other hand, this paper focuses on BLI scores only, which were shown to not correlate perfectly with bilingual performance on downstream tasks (Glavaš et al., 2019). We suppose that to some extent our findings hold in downstream situations as well, since downstream performance cannot be independent of BLI scores, but this question

L-Wiki:			
	VecMap	**RCSLS**	**NAT**
Laplacian	0.737	0.740	0.754
SVG	0.617	0.595	0.601
Spectral	0.760	0.777	0.770
Effective Spectral	0.763	0.766	0.776
Relational	0.735	0.727	0.731
Neighbors	0.793	0.796	0.802
M-Wiki:			
	VecMap	**RCSLS**	**NAT**
Laplacian	0.524	0.517	0.537
SVG	0.746	0.731	0.737
Spectral	0.821	0.827	0.824
Effective Spectral	0.815	0.832	0.824
Relational	0.681	0.678	0.660
Neighbors	0.553	0.570	0.564
S-Wiki:			
	VecMap	**RCSLS**	**NAT**
Laplacian	0.443	0.453	0.437
SVG	0.621	0.655	0.649
Spectral	0.448	0.456	0.448
Effective Spectral	0.537	0.552	0.548
Relational	0.746	0.724	0.743
Neighbors	0.642	0.655	0.645

Table 5: R2 scores of the predictions of isomorphism, based on our three corpus properties combined, using random forest regression.

should be part of further research. The main difference between downstream and BLI evaluation scores is probably the importance of monolingual embedding quality: while two embedding matrices can be trained almost perfectly isomorphically on a relatively small parallel corpus, the monolingual performance of these embeddings probably lags behind embeddings trained on a big corpus, Wikipedia for example. But at the same time, this also shows that embeddings can be mapped very well even if they are not of the highest quality, but their corpora are similar enough.

Acknowledgements

Réka Cserháti was supported by the ÚNKP-21-1 – New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund.

The research presented in this paper was partly supported by the Ministry of Innovation and the National Research, Development and Innovation Office within the framework of the Artificial Intelligence National Laboratory Programme.

References

Simran Arora, Avner May, Jian Zhang, and Christopher Ré. 2020. Contextual embeddings: When are they worth it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2663, Online. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. 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)*, pages 789–798.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Yo Joong Choe, Kyubyong Park, and Dongwoo Kim. 2020. word2word: A collection of bilingual lexicons for 3,564 language pairs. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3036–3045, Marseille, France. European Language Resources Association.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herv'e J'egou. 2018. Word translation without parallel data. *ArXiv*, abs/1710.04087.

Yerai Doval, Jose Camacho-Collados, Luis Espinosa Anke, and Steven Schockaert. 2020. On the robustness of unsupervised and semi-supervised cross-lingual word embedding learning. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4013–4023, Marseille, France. European Language Resources Association.

Haim Dubossarsky, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2020. The secret is in the spectra: Predicting cross-lingual task performance with spectral similarity measures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2377–2390, Online. Association for Computational Linguistics.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 469–478, Brussels, Belgium. Association for Computational Linguistics.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. 2019. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3336–3341, Hong Kong, China. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Robert Litschko, Ivan Vulic, Simone Paolo Ponzetto, and Goran Glavas. 2021. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I*, volume 12656 of *Lecture Notes in Computer Science*, pages 342–358. Springer.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Neural and Information Processing System (NIPS)*.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4990–4995, Florence, Italy. Association for Computational Linguistics.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of*

the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4407–4418, Hong Kong, China. Association for Computational Linguistics.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online. Association for Computational Linguistics.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.