Unsupervised Geometric and Topological Approaches for Cross-Lingual Sentence Representation and Comparison

Shaked Haim Meirom[1] and **Omer Bobrowski**[1,2]
[1] Viterbi Faculty of Electrical & Computer Engineering, Technion - Israel Institute of Technology
[2] School of Mathematical Sciences, Queen Mary University of London
{shakedmeirom@campus, omer@ee}.technion.ac.il

Abstract

We propose novel structural-based approaches for the generation and comparison of cross lingual sentence representations. We do so by applying geometric and topological methods to analyze the structure of sentences, as captured by their word embeddings. The key properties of our methods are: (a) They are designed to be *isometric invariant*, in order to provide language-agnostic representations. (b) They are fully unsupervised, and use no cross-lingual signal. The quality of our representations, and their preservation across languages, are evaluated in similarity comparison tasks, achieving competitive results. Furthermore, we show that our structural-based representations can be combined with existing methods for improved results.

1 Introduction

Word embeddings are driven by distributional concepts, i.e. words can be described by their surrounding words. For example, the words "dog" and "cat" often occur in similar contexts, and therefore their embeddings are expected to be nearby, and to have similar distances to other words. Such similarities are inherent to the real world (e.g., a dog is similar to a cat), and therefore should be language-agnostic. For that reason, the embedding spaces of different languages are expected to be near-isomorphic (Miceli Barone, 2016). Notably, Vulić et al. (2020) demonstrated that high degree of isomorphism can be reached with sufficient monolingual resources. This assumption has enabled various applications at the word level, e.g. generating cross-lingual word embeddings by mapping monolingual vector spaces (Artetxe et al., 2018; Conneau et al., 2018).

In this paper, we take a step further, leveraging the approximate isomorphism between monolingual spaces at the *sentence* level. Considering each sentence as a point cloud (made by its word embeddings), our key argument is that these point clouds retain geometric and topological structures that should be preserved across languages. Therefore, they have the potential to enable language-agnostic sentence representations.

We investigate different approaches for extracting and utilizing such structures. Firstly, we devise a geometric approach, based on the intra-distances of the word embeddings in a sentence. Secondly, we explore a topological approach, borrowing methods from Topological Data Analysis (TDA). Briefly, TDA provides algebraic-topological methods to extract global structural information from shapes. These methods are coordinate free and invariant to isometries (Carlsson, 2009; Zomorodian, 2012), which is highly desired in our setting. Our main goal is to employ these structure-based features to generate novel cross-lingual sentence representations, in a *fully unsupervised* manner.

In order to evaluate the cross-lingual nature of our representations, we experiment with similarity comparison tasks, including bilingual sentence retrieval (Guo et al., 2018) and machine translation quality estimation (Specia et al., 2020).

Our contributions can be summarized as follows. (1) Proposing the novel concept of exploiting the isomorphism of word embedding spaces at the sentence level. (2) Devising fully unsupervised methods for cross-lingual sentence representation, based on geometric and topological approaches. (3) Providing measures of similarity for the new representations. (4) Evaluating the extent to which these representations are preserved across languages, via downstream similarity comparison tasks.

2 Isometry of Word Embedding Spaces

Measuring and utilizing the similarities between word embedding spaces, is a well-studied topic in NLP. In this context, a standard assumption is that monolingual word embedding spaces are approximately isomorphic. A common use for such near-isomorphism is to search for a linear trans-

Proceedings of the 7th Workshop on Representation Learning for NLP, pages 173 - 183
May 26, 2022 ©2022 Association for Computational Linguistics

formation between the embedding spaces of different languages (Artetxe et al., 2018; Mikolov et al., 2013a; Glavaš et al., 2019). Other studies argue that a better practice is to consider orthogonal transformations (Xing et al., 2015; Smith et al., 2017). These transformations have been used to induce bilingual dictionaries (Xing et al., 2015; Artetxe et al., 2018), as well as cross-lingual transfer learning (Ruder et al., 2019). In fact, mapping-based approaches have become a prevalent way to learn cross-lingual embedding spaces.

The isomorphism assumption is also used in fully unsupervised settings, including unsupervised bilingual lexicon induction (Artetxe et al., 2018; Conneau et al., 2018) and unsupervised machine translation (Lample et al., 2018; Artetxe et al., 2019). Here, the alignment between the monolingual embedding spaces cannot be achieved by mapping pre-existing bilingual dictionaries. Instead, it is achieved either by using adversarial training (Conneau et al., 2018) or by comparing the distribution of similarities or distances of the word embeddings across languages (Artetxe et al., 2018; Alvarez-Melis and Jaakkola, 2018).

As explained by Xu and Koehn (2021), the isomorphism of embedding spaces can be extended to isometry, using normalization techniques. As the isometry of word embedding spaces becomes the premise for a large variety of methods, the following question arises: **can we leverage the isometry of word embeddings at the *sentence* level?**

Our approach is to take the embedding of a sentence to be the word-by-word embedding, resulting in a point-cloud (finite collection of points). Assuming nearly-isometric word embeddings, one would expect that the geometric and topological structures of these point clouds are preserved across languages to some extent. For this reason, we devised methods to extract such structural information from sentences and provide means to compare the structures of different sentences.

3 Related Work

Various studies of unsupervised cross-lingual sentence representations rely on aggregation of either mapped word embeddings or contextualized word embeddings from pre-trained multilingual models (Smith et al., 2017; Conneau et al., 2018; Xu and Koehn, 2021; Kvapilíková et al., 2020). These studies often rely (implicitly or explicitly) on the isometry assumption between word vector spaces, for constructing cross lingual mappings. However, they do not utilize the isometry for generating sentence representations.

Closer to our work are studies using structural similarities between languages. Both Aldarmaki et al. (2018) and Alvarez-Melis and Jaakkola (2018) exploit the preservation of geometric structures between monolingual vector spaces for cross lingual mapping of word embeddings. However, neither refer to geometric structures of sentences.

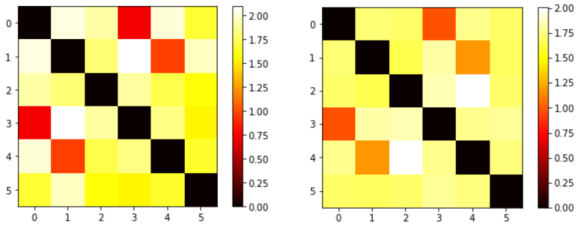
Finally, several studies have used topological approaches in NLP related tasks, such as word sense disambiguation (Jakubowski et al., 2020) and text visualization (Sami and Farrahi, 2017). TDA methods were also used to generate sentence and document representations. Zhu (2013) was the first to introduce the concept of topological text representation. Built on this idea, recent studies designed various methods for document representations by computing persistent homology (see Section 6.2) over their word embeddings. These methods were evaluated on tasks such as document classification and discourse analysis (Tymochko et al., 2020; Gholizadeh et al., 2020; Savle et al., 2019). Most related to our work is Michel et al. (2017), where persistence diagrams were used to represent documents and sentences. Their final representation and comparison of sentences are quite different than ours, and achieved negative results in classification and clustering tasks. To the best of our knowledge, no previous study used topological-based approaches in cross-lingual tasks.

4 Sentence Distance Matrix

In this section we present the fundamental element of our pipeline – the *Sentence Distance Matrix* (SDM). Representing sentences by point clouds, the geometric information about the sentence can be encoded by the pairwise distances between the words. Formally, let $X = (x_1, ..., x_n)$ be a collection of word embeddings. We define SDM_X to be the $n \times n$ matrix whose entries are given by $(\mathrm{SDM}_X)_{i,j} = \mathrm{dist}(x_i, x_j)$, where dist can be any metric in the embedding space.

The motivation for using SDMs, is that in the hypothetical case where X and Y represent equal-length sentences[1], with parallel words, and in languages with perfectly-isometric word embeddings, we have $\mathrm{SDM}_X = \mathrm{SDM}_Y$. Realistically, while translated words are not always parallel, they are ex-

[1]We will treat non-equal sentence lengths in Section 6.1.



(a) ‘dog’, ‘stood’, ‘floor’, ‘cat’, ‘sat’, ‘mat’ (b) ‘perro’, ‘paró’, ‘suelo’, ‘gato’ , ‘sentó’, ‘alfombra’

Figure 1: SDMs for an English sentence (a) and its Spanish translation (b), stopwords removed. Note the pairs dog-cat and stood-sat are close in both languages.

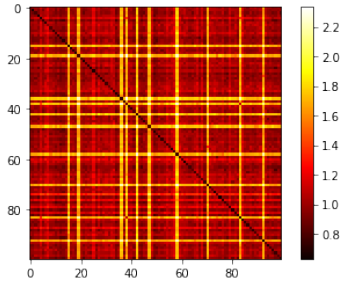


Figure 2: Distances between SDMs of English sentences and SDMs of parallel Italian sentences.

Accuracy	En-It	It-En
P@1	0.939 ± 0.008	0.902 ± 0.006
P@5	0.979 ± 0.003	0.957 ± 0.006
P@10	0.987 ± 0.003	0.969 ± 0.05

Table 1: Results of bilingual sentence retrieval, based on word-by-word translations, using SDMs as the distance measure between sentences.

pected to be semantically related. In addition, while perfect-isometry does not exist, we do expect to have near-isometric embeddings. Thus, we expect translated sentences to have $\mathrm{SDM}_X \approx \mathrm{SDM}_Y$, implying that the SDM is a good candidate to represent and compare the structure of sentences.

We demonstrate the resemblance between the SDMs of sentences in different languages in the following example. The Spanish sentence:

> *El perro se paró en el suelo y el gato se sentó en la alfombra*,

is a translation of the following English sentence:

> *The dog stood on the floor and the cat sat on the mat*.

The SDMs of these sentences are presented in Figure 1. The resemblance between the English sentence and its Spanish translation is apparent through their SDMs.

Defining suitable metrics to compare between SDMs (see Sections 5 and 6), will enable us measure similarity between sentences in different languages, *without any supervised or bilingual signal*. This measurement can be useful in many NLP tasks, such as Machine Translation Quality Estimation (Specia et al., 2020), Parallel Corpus Filtering (Koehn et al., 2020), Parallel Corpus mining (Guo et al., 2018) and Cross-Lingual Plagiarism Detection (Danilova, 2013). In addition, such multilingual representations can be useful in cross lingual transfer learning (Ruder et al., 2019).

5 Toy Example

To demonstrate the potential of SDMs, we start with a simple experiment. As we argued earlier, the SDMs of sentences and their translations should be similar, especially if the translation uses parallel words (word-by-word translation). This suggests that SDMs can achieve high performance in a suitable bilingual sentence retrieval setting, which measures the accuracy on retrieving the translation of a sentence from a given bilingual corpus.

The experiment settings are as follows. We use the English-Italian dataset provided by Dinu et al. (2014) and Artetxe et al. (2016). The dataset contains monolingual word embeddings trained with word2vec using the CBOW method with negative sampling (Mikolov et al., 2013a)[2]. We apply length normalization and mean centering to all embeddings. In addition to the embeddings, the dataset also contains a bilingual dictionary, split into a training set of 5,000 word pairs and a test set of 1,500 word pairs, both are uniformly distributed in frequency bins. We use the training bilingual dictionary, and removed repetitions[3], which resulted in a dictionary of 3,281 word pairs. We generate a bilingual corpus of 1,000 artificial sentence pairs using random sampling from the bilingual dictionary, such that each sentence is made of a sequence of 20 random words, and its parallel sentence is made of the translations of these words.

Next, we calculate the SDM of each sentence (a 20×20 distance matrix), using the Euclidean distance. In order to measure the distance between the English and Italian sentences, we use the Frobenius norm of the difference between their respective SDMs. This results in a 1000×1000 distance matrix, the first 100×100 block of which is presented

[2]The hyper-parameters and corpora used to create the dataset are described in Artetxe et al. (2016).

[3]Words which appear more than once in the dictionary.

in Figure 2. The sentences have the same order in both languages. Therefore, the distance between each sentence and its translation appears in the diagonal. One can easily notice that the diagonal tends to contain the lowest values, supporting our intuition that translations should have the closest SDM to their source sentences.

To provide quantitative evaluation we used the mean accuracy measure. We count how many times the correct translation of a source sentence is retrieved, and report mean precision@k for $k = 1, 5, 10$, by repeating the experiment 10 times. The results are provided in Table 1. Note that even though the SDMs do not rely on any supervised or bilingual signal, the results are near-perfect. This demonstrates the potential of SDMs in cross-lingual settings, and the preservation of the geometric structure of sentences across languages.

6 Methods

In this section, we describe how to utilize SDMs for sentence representation in the realistic case, where parallel sentences in different languages may differ in length and word ordering. Section 6.1 describes a direct approach – interpolating the SDMs of parallel sentences to have the same size, enabling a direct matrix comparison. Section 6.2 describes a vastly different approach – extracting topological structure information from the SDMs using TDA.

6.1 A Geometric Approach

In Section 5 we showed that the Frobenius norm is an effective method to measure similarity between SDMs of sentences in different languages. However, this procedure requires that the compared sentences share the same length and order. In reality however, this is often not the case. In this section we propose a framework that generalizes this procedure to the most generic setting.

The first challenge to address is different sentence lengths. To this end, we propose to rescale the SDM matrices. Matrix rescaling is a fundamental challenge in the field of image processing, e.g. when zooming in (upscaling) or zooming out (downscaling). We use the well-known B-spline interpolation method introduced by Hou and Andrews (1978), and refined by Unser et al. (1991). Briefly, in order to upscale an SDM we find the piecewise polynomial function that best approximates the original matrix values, and then sample this function at the desired resolution. The resulting pipeline is presented in Figure 3, and is referred to as *interpolated SDM* (ISDM).

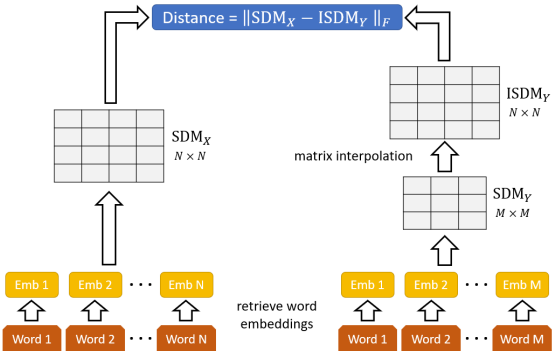


Figure 3: Interpolated SDMs pipeline. The word embeddings of each sentence are used to generate an SDM. The smaller SDM is then interpolated to match the size of the larger one. Using Frobenius norm we can now measure the distance between the sentences.

The next challenge we need to address is the different word ordering between parallel sentences. Intuitively, we propose to match words based on their geometric representation (encoded by the SDM) rather than their position within the sentence, without any bilingual signal. Given two sentences, we take the columns of their interpolated SDMs (vaguely representing words) and search for the optimal matching that minimizes the Frobenius norm. This variation of the pipeline is referred to as *order-aware interpolated SDM* (OSDM).

6.2 A Topological Approach

In this section, we propose a vastly different method to represent and compare sentences, by extracting robust information from the SDMs, describing the topological structure of sentences.

6.2.1 Topological Data Analysis

Topological Data Analysis (TDA) promotes the use of mathematical topology in analyzing data and networks (Carlsson, 2009; Zomorodian, 2012; Zhu, 2013). The key idea is that topology can be used to study the shape of data in a qualitative way that is isometric invariant and robust to continuous deformations. In this section we briefly introduce the relevant concepts and tools of TDA, and discuss how to adapt them for sentence analysis.

The Vietoris-Rips complex. A *simplicial complex* is a high-dimensional generalization of a graph, consisting of vertices, edges, triangles, tetrahedra, and higher dimensional faces. In order to extract structural information from point clouds (word embeddings in our setting), a common practice in TDA

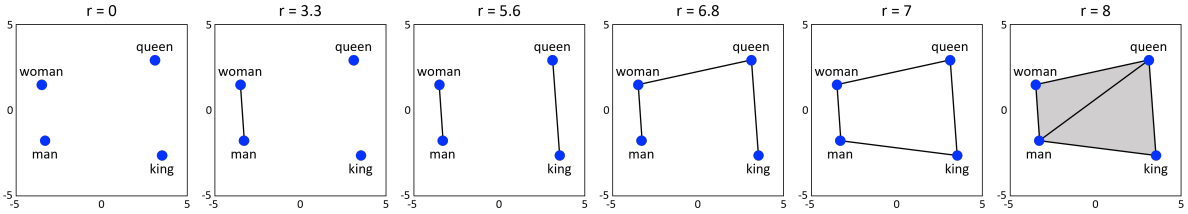


Figure 4: VR complexes of word embeddings, for an increasing diameter r (using the Euclidean distance). For instance, $\mathrm{VR}_{(r=5.6)}$ includes two 1-dimensional faces (edges), since there are two subsets of size 2, whose diameter is less than 5.6. The embeddings were extracted using GloVe (Pennington et al., 2014), and transformed to $\mathbb{R}^2$ using PCA, for visualization purposes. Note that the last step introduces two 2-dimensional faces (triangles).

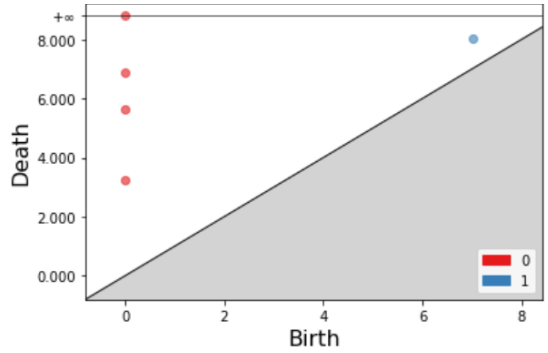


Figure 5: Persistence diagram for the sequence of VR complexes in Figure 4. The red points mark connected components (0-cycles), and their deaths occur when two components merge. For example, at $r \approx 3.3$ a 0-dimensional cycle dies, as the components of "man" and "woman" merge. The blue point marks a hole (1-cycle), appearing at $r \approx 7$ and later filled in at $r \approx 8$.

is to first construct a simplicial complex known as the Vietoris-Rips (VR) complex. Given a point cloud P, the vertex set of $\mathrm{VR}_r(P)$ is just P, and its k-dimensional faces are all subsets $S \subset P$ of size $k+1$, whose diameter is less than r. In Figure 4 we present a sequence of VR complexes, for a point cloud P made by the word embeddings of the well-known word set "king", "queen", "man", and "woman".

Homology is an algebraic topological structure that characterizes the shape of topological spaces. If X is a topological space (e.g., the VR complex), we attach to it a sequence of vector spaces (or groups) denoted $H_0(X)$, $H_1(X)$, $H_2(X)$, etc. The basis elements of $H_0(X)$ correspond to the connected components (referred to as 0-cycles) of X, $H_1(X)$ – to loops surrounding holes in X (1-cycles), $H_2(X)$ – to closed surfaces enclosing "bubbles" in X (2-cycles). Generally, $H_k(X)$ represents information about "k-dimensional cycles", which can be thought of as k-dimensional surfaces that are empty from within. For more details, see (Hatcher, 2002).

Persistent Homology (PH) is the core method used in TDA, whose goal is to extract robust multi-scale topological information from data. Consider the $\mathrm{VR}_r(P)$ complex described above. Increasing the value of r, k-cycles may form at various times (r), and later terminate (merge with another component or fill in). The k-th persistent homology, denoted PH_k, tracks this birth-death process. The information provided by PH_k is often summarized by a *persistence diagram*, which is a collection of points in the plane, where the x and y coordinates represent the birth and death times of a cycle, respectively. In Figure 5 we present the persistence diagram extracted from the sequence of VR complexes in Figure 4. This example demonstrates the unique information captured by PH, which in this example highlights the circular relationship between the words king→queen→woman→man→king.

Wasserstein Distance is the most commonly used metric to compare between persistence diagrams, based on an optimal matchings of their points. For every two diagrams D_1, D_2 we denote by $\hat{D}_1, \hat{D}_2$ their augmented versions that include the diagonal line (death=birth). This allow for matchings that add or remove points from each diagram, by assigning them with the nearest point on the diagonal. The p-Wasserstein distance is then

$$W_p(D_1, D_2) := \inf_{\phi:\hat{D}_1 \to \hat{D}_2} \Big(\sum_{x \in \hat{D}_1} \|x - \phi(x)\|^p \Big)^{1/p},$$

where ϕ goes over all possible bijections.

6.2.2 Order-Aware Persistence Diagrams

We wish to use persistent homology to extract and compare the structural information of sentences. In order to do so, we take our point clouds to be the word embeddings of a sentence, and compute the persistent homology for the VR complex, using the distances calculated by the SDM. To measure similaity between two sentences, we use the Wasser-

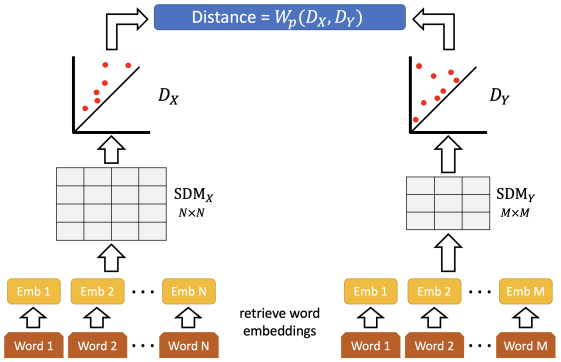


Figure 6: Topological distance pipeline. The word embeddings of each sentence are used to generate an SDM. The SDMs are used to generate the persistence diagrams for the VR complex. The Wasserstein distance is used to measure the similarity between the diagrams.

stein distance of their corresponding persistence diagrams. See Figure 6 for the complete pipeline.

Note that each step of the pipeline is isometric invariant, and therefore allows for cross-lingual representation and similarity measurement. In addition, while the proposed method compares structural information of sentences, it does *not* require them to have the same length (as opposed to the comparison of raw SDMs). One drawback of this pipeline is that it is oblivious to the word ordering within sentences. Next, we wish to present two possible solutions to address this issue.

The first solution is to enrich the diagrams with the positions of the words within a given sentence. Note that in the case of the VR complex, each death event in PH corresponds to an edge e^* entering the complex, as the parameter value r increases. Denote by $\mathrm{pos}_1, \mathrm{pos}_2$ the positions of the words that are the end-points of e^*. We create an augmented diagram, where each cycle is represented by four coordinates $(\mathrm{birth}, \mathrm{death}, \mathrm{pos}_1, \mathrm{pos}_2)$. We refer to the result as the *order-aware persistence diagram* (OPD). To compare between two such diagrams, D_1 and D_2, we devised an adaptation of the Wasserstein distance, where the diagonal of D_1 is augmented by taking the $\mathrm{pos}_1, \mathrm{pos}_2$ values to be the average word positions of all cycles in D_2, and vice verse.

The second method is inspired by the "time skeleton" concept suggested by Zhu (2013). The key idea is to encode the flow of the sentence into the VR complex, by placing an edge between every two words at adjacent positions in the sentence (at $r = 0$), independently of the distance between their embeddings. We refer to these edges as *sequence edges*. This method increases the expressiveness of H_1 (i.e., holes), since all adjacent words are connected immediately, enabling early appearances of holes. On the other hand, note that the resulting VR complex is always connected, hence H_0 is trivial.

7 Experiments and Results

In this section we want to examine the preservation of our new representations across languages, by evaluating their performance in real-world tasks. In particular, we will focus on tasks that are based on similarity between parallel sentences. Note that the proposed methods are *fully unsupervised*, in the sense that they do not use any task-specific training or cross-lingual data. We will evaluate the effectiveness of our methods as well as their combination with other unsupervised methods (i.e. without increasing the level of supervision).

7.1 Bilingual Sentence Retrieval

The objective of bilingual sentence retrieval is to find the translation of sentences in a source language from a list of candidates in the target language. In this section we want to show that our methods can be used to enhance the performance of existing semantic-based methods for the fully unsupervised version of this task.

We evaluate on the English-Spanish and English-Russian language pairs of the UN parallel corpus (Ziemski et al., 2016). We consider 2,000 source sentences queries and 20,000 possible target sentences for each direction[4]. For the monolingual word representations, we use pre-trained fasttext word embeddings (Grave et al., 2018).

For the baseline, we use the fully unsupervised version of Vecmap (Artetxe et al., 2018) to map monolingual word embeddings into a cross-lingual space. We aggregate the word embeddings by mean-pooling, in order to represent sentences. The pipeline we examine has two steps: (1) use the baseline to list the top 10 nearest neighbours of each source sentence. (2) Re-rank this list using our novel representations.

For the first step, we score all possible sentence pairs using the cosine distance between their embeddings. We mitigate the hubness problem of embedding spaces using the margin-based approach of Artetxe and Schwenk (2019)[5]. We then create

[4]We considered sentences with at least 5 words, and stripped punctuation as a pre-processing step.

[5]We used the Ratio variant with parameter $k = 10$.

	English-Spanish						English-Russian					
	En→Es			Es→En			En→Ru			Ru→En		
	P@1	**P@5**	**MAP**	**P@1**	**P@5**	**MAP**	**P@1**	**P@5**	**MAP**	**P@1**	**P@5**	**MAP**
Vecmap	45.4	64.0	.535	56.2	73.3	.634	36.3	55.9	.448	45.1	65.0	.535
ISDM	39.4	58.0	.480	56.6	71.7	.634	37.1	55.6	.449	42.5	61.1	.511
OSDM	40.3	60.1	.490	4.93	70.9	.586	**42.7**	**58.7**	**.494**	35.2	57.7	.452
OPD_0	52.7	68.5	.593	60.5	76.1	.671	40.5	58.2	.480	51.1	69.7	.589
OPD_1	51.5	67.3	.583	59.7	76.1	.665	39.2	58.0	.469	50.8	68.4	.583
OPD_{0+1}	**53.6**	**68.5**	**.599**	**60.7**	**76.4**	**.672**	40.6	58.4	.480	**51.5**	**69.7**	**.592**
OPD_2	46.7	64.8	.545	57.0	74.1	.641	38.1	56.6	.461	46.1	66.6	.546
OPD_{0+1+2}	51.7	67.2	.584	58.9	74.8	.657	40.3	59.1	.479	49.7	68.7	.578

Table 2: Results for the fully unsupervised bilingual sentence retrieval, as described in Section 7.1. OPD_{a+b} stands for a linear combination between the baseline, OPD_a and OPD_b. We highlight the best result for each direction.

an ordered list of the top 10 nearest neighbours of each source sentence.

In the second step, we wish to enhance the baseline ranking by applying our new geometric and topological methods. The methods we examine are: (1) interpolated SDM (ISDM), (2) order-aware interpolated SDM (OSDM), and (3) order-aware k-cycles persistence diagrams (OPD_k)[6]. We use each of these methods to compute the distance between every source sentence and its 10 nearest neighbors, found in step 1. We note that the calculations in this step are applied directly to the monolingual word embeddings (rather than the Vecmap embeddings). Next, we create new scores for each sentence pair by a linear combination of the baseline distance (from step 1) and the structure-based distances[7]. Finally, we re-rank the top nearest neighbours lists according to the new scores. As this is a retrieval task, we follow Glavaš et al. (2019) and use the Mean Average Precision (MAP), in addition to precision@k (with $k \in \{1, 5\}$) for the evaluation.

We report the average results (across all sentence queries) in Table 2. As can be seen, using our structure-based methods improves the results of the baseline on all fronts. Remarkably, the improvement (15% on average for P@1, and 10% on average for MAP), does not rely on any additional data or training. In most cases, the combination between OPD_0 and OPD_1 yields the best results, except for one case in which the OSDM triumphs. It is also interesting to note that the distance provided by OPD_2 improves the results as well. While the structural information provided by 2-cycles is less intuitive (and consequently is uncommonly used in applications), our results indicate that such high-dimensional topological structures do carry significant information in language processing.

7.2 Machine Translation Quality Estimation

The goal here is to predict quality scores for translated sentences, in a way that is consistent with human perceived scores, referred to as *direct assessment*. Since the objective is to compare parallel sentences, this is a suitable scenario to test our novel representations across languages.

The implementation details are as follows. We generate monolingual word representations, based on pre-trained BPEmb subword embeddings (Heinzerling and Strube, 2018). These embeddings were chosen in order to properly deal with out-of-vocabulary words. For words that consist of multiple subwords, we take average of the subword vectors. This common practice outperforms other aggregation methods (Bommasani et al., 2020).

We use the monolingual embeddings to calculate the structural-based distances between every source sentence and its translation[8], using the methods proposed in Section 6. We take the inverse of the distance as the predicted quality score. We tested our methods separately as well as combined (taking linear combinations of the respective scores[9]). We note that for the topological approach, we always used OPD_0 and OPD_1 together, as this method demonstrated superior results.

We tested this pipeline on the language pairs English-German (en-de) and Sinhala-English (si-en), of the WMT2020 Quality Estimation shared task (Specia et al., 2020). Each language pair includes 1,000 source sentences and their translations, produced by state-of-the-art NMT models.

[6]For the OPD_1 we utilize the sequence edges method.

[7]The weights of the linear combination were chosen according to preliminary experiments, and were usually balanced, slightly favoring our methods.

[8]As a pre-processing step, we remove stopwords and stripped punctuation.

[9]The coefficients were optimized manually in preliminary experiments.

We compare the results of our methods to the supervised baseline of the shared task, which uses LSTM-based Predictor-Estimator approach (Kim et al., 2017), and to the following competitors.

TransQuest (Ranasinghe et al., 2020) is the winner method of the shared task. The method uses an ensemble of two architectures, which rely on pre-trained XLM-R large transformer models, and are fine-tuned on quality estimation datasets.

FVCRC (Zhou et al., 2020) is an unsupervised method based on a BERTScore (Zhang et al., 2020). The method relies on pre-trained transformer-based models (mBert, XLM) to extract word (or subword) embeddings. It aligns the embeddings using cosine similarity based greedy matching, and predicts the quality score as the sum of the respective similarities. It enhances the alignments using explicit cross-lingual knowledge from external models.

Bergamot-LATTE, glass-box (Fomicheva et al., 2020) is an unsupervised method that assumes access to the machine translation model. It extracts features from the model output and uses uncertainty quantification to predict the translation quality.

As most of the competing methods rely on pre-trained transformer-based models, we also wish to evaluate the merge between these models and our framework. We do so in a way that keeps the combined pipeline fully unsupervised (avoiding fine-tuning and bilingual knowledge). To this end, we adapted the FVCRC approach to be fully unsupervised, replacing their alignment procedure with optimal transportation matching[10]. We refer to this fully unsupervised transformer-based method as *cross lingual matching* (CLM)[11].

Following the shared task guidelines, we present the Pearson correlation between the predicted and the manually annotated scores in Table 3. Note that both of our approaches (geometric and topological) provide meaningful and competitive results, even though they are *fully unsupervised* and do not rely on any cross-lingual signal or external model. Interestingly, the combinations between our geometric and topological approaches has yielded superior results. More specifically, the results reveal that our methods outperform the *supervised* baseline as well as the unsupervised methods in the en-de direction. In addition, in the si-en direction, the combination between CLM and our methods performs better than its competitor – the unsupervised FVCRC, and better than the supervised baseline.

[10] Specifically we use the Sinkhorn distance (Cuturi, 2013).

[11] As FVCRC, our implementation also uses BERTScore.

Method	Supervision	en→de	si→en
TransQuest	Sup.	0.55	0.68
Baseline		0.15	0.37
FVCRC	Unsup.*	0.11	**0.39**
Bergamot-LATTE	Unsup.**	**0.26**	**0.51**
ISDM	Fully Unsup.	0.12	0.26
OSDM		**0.27**	0.16
OPD		**0.20**	0.19
ISDM + OPD		**0.19**	0.29
OSDM + OPD		**0.27**	0.18
CLM + OPD		0.13	**0.45**
CLM + OSDM		0.14	**0.45**

Table 3: Pearson correlation with direct assessment scores for the WMT2020 Machine Translation Quality Estimation shared task. The '+' sign stands for a linear combination between methods. Unsupervised results improving the supervised baseline are highlighted. *FVCRC uses explicit bilingual signal. **Bergamot-LATTE relies on the MT model scores.

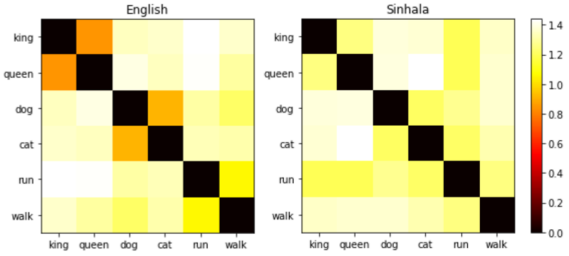


Figure 7: Comparing word embeddings of English and Sinhala. To demonstrate the inferior quality of the Sinhala word embeddings, we present the distance matrices for six words in English and their translations in Sinhala. We observe that similar meaning corresponds to short distances between the word embeddings in English. However, the same is not true for Sinhala.

We attribute the relative lower performance of our methods in the si-en direction to the poor representation of the monolingual word embeddings for Sinhala, as a low resource language. The representation capability and the expressiveness of the distances between the pre-trained word embeddings are demonstrated at Figure 7. Generally, the performance of our methods will be improved if the degree of isomorphism between the relevant word vector spaces is increased. This can be achieved, for example, by training the word embeddings with additional monolingual data, as suggested by Vulić et al. (2020). This is left as future work.

8 Conclusion and Future Work

We introduced the concept of leveraging the isometry of word embedding spaces at the sentence

level. This enabled us to propose geometric and topological approaches that facilitate fully unsupervised generation of cross-lingual sentence representations, together with suitable similarity measures.

We conducted cross-lingual experiments, where our standalone methods have achieved competitive results on different tasks. Moreover, we observed that combining our methods with traditional ones has led to notable enhanced performance. We should emphasize that this was achieved *without any additional data or training*. We conclude that geometric and topological structures of sentences are preserved to a significant level across languages. Interestingly, our experiments show that the shapes we extract have complex structures. For example, in many cases we found meaningful homological cycles in various degrees. We note that these representations are weaker on scenarios with low degree of isomorphism, e.g. due to lack of monolingual data (Vulić et al., 2020).

A promising direction for future work is to utilize the proposed representations in cross-lingual transfer learning (training a model on one language and using it on another language).

Finally, we note that the main motivation for this work was to promote the use of geometric and topological approaches in core NLP tasks, and especially cross-lingual tasks. We believe that the ideas and methods we presented here will contribute to the future development of this line of research.

Acknowledgements

The authors are grateful to Roi Reichart for helpful comments and feedback, especially regarding the relevant NLP tasks. OB was supported in part by the Israel Science Foundation, Grant 1965/19.

References

Hanan Aldarmaki, Mahesh Mohan, and Mona Diab. 2018. Unsupervised word mapping using structural similarities in monolingual embeddings. *Transactions of the Association for Computational Linguistics*, 6:185–196.

David Alvarez-Melis and Tommi Jaakkola. 2018. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890.

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.

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781.

Gunnar Carlsson. 2009. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. *Proceedings of ICLR 2018*.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pages 2292–2300. Curran Associates, Inc.

Vera Danilova. 2013. Cross-language plagiarism detection methods. In *Proceedings of the Student Research Workshop associated with RANLP 2013*, pages 51–57.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *Proceedings of the 3rd International Conference on Learning Representations (ICLR2015), workshop track*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Vishrav Chaudhary, Mark Fishel, Francisco Guzmán, and Lucia Specia. 2020. BERGAMOT-LATTE submissions for the WMT20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1010–1017.

Shafie Gholizadeh, Armin Seyeditabari, and Wlodek Zadrozny. 2020. A novel method of extracting topological features from word embeddings. *arXiv preprint arXiv:2003.13074*.

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.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. "effective parallel corpus mining using bilingual sentence embeddings". In *"Proceedings of the Third Conference on Machine Translation: Research Papers"*, pages "165–176".

Allen Hatcher. 2002. *Algebraic topology*. Cambridge University Press.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Hsieh Hou and H Andrews. 1978. Cubic splines for image interpolation and digital filtering. *IEEE Transactions on acoustics, speech, and signal processing*, 26(6):508–517.

Alexander Jakubowski, Milica Gasic, and Marcus Zibrowius. 2020. Topology of word embeddings: Singularities reflect polysemy. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 103–113.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the wmt 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar. 2020. Unsupervised multilingual sentence embeddings for parallel corpus mining. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 255–262.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 121–126.

Paul Michel, Abhilasha Ravichander, and Shruti Rijhwani. 2017. Does the geometry of word embeddings help document classification? a case study on persistent homology-based representations. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 235–240.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, page "3111–3119".

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, page 569–630.

Ishrat Rahman Sami and Katayoun Farrahi. 2017. A simplified topological representation of text for local and global context. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1451–1456.

Ketki Savle, Wlodek Zadrozny, and Minwoo Lee. 2019. Topological data analysis for discourse semantics? In *Proceedings of the 13th International Conference on Computational Semantics - Student Papers*, pages 34–43.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *International Conference on Learning Representations*.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Sarah Tymochko, Zachary New, Lucius Bynum, Emilie Purvine, Timothy Doster, Julien Chaput, and Tegan Emerson. 2020. Argumentative topology: Finding loop(holes) in logic. *arXiv preprint arXiv:2011.08952*.

Michael Unser, Akram Aldroubi, Murray Eden, et al. 1991. Fast b-spline transforms for continuous image representation and interpolation. *IEEE Transactions on pattern analysis and machine intelligence*, 13(3):277–285.

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

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.

Haoran Xu and Philipp Koehn. 2021. Cross-lingual bert contextual embedding space mapping with isotropic and isometric conditions. *arXiv preprint arXiv:2107.09186*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lei Zhou, Liang Ding, and Koichi Takeda. 2020. Zero-shot translation quality estimation with explicit cross-lingual patterns. *Proceedings of the 5th Conference on Machine Translation (WMT)*, page 1068–1074.

Xiaojin Zhu. 2013. Persistent homology: An introduction and a new text representation for natural language processing. In *IJCAI*, pages 1953–1959.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.

Afra Zomorodian. 2012. Topological data analysis. *Advances in applied and computational topology*, 70:1–39.