# Handling the Deviation from Isometry Between Domains and Languages in Word Embeddings: Applications to Biomedical Text Translation

Félix Gaschi, Parisa Rastin, Yannick Toussaint

# Handling the Deviation from Isometry between Domains and Languages in Word Embeddings: Applications to Biomedical Text Translation

Félix Gaschi[1,2] ✉, Parisa Rastin[2], and Yannick Toussaint[2]

[1] SAS Posos, 55 rue de la Boétie, 75008 Paris, France
felix@posos.fr
[2] LORIA, UMR 7503, BP 239, 54506 Vandoeuvre-lès-Nancy, France
{felix.gaschi,parisa.rastin,yannick.toussaint}@loria.fr

**Abstract.** Previous literature has shown that it is possible to align word embeddings from different languages with unsupervised methods based on a distance-preserving mapping, with the assumption that the embeddings are isometric. However, these methods seem to work only when both embeddings are trained on the same domain. Nonetheless, we hypothesize that the deviation from isometry might be reduced between relevant subsets of embeddings from different domains, which would allow to partially align them. To support our hypothesis, we leverage the Bottleneck distance, a topological data analysis tool used to approximate the deviation from isometry. We also propose a cross-domain and cross-lingual unsupervised alignment method based on a proxy embedding, as a first step towards new cross-lingual alignment methods that generalize to different domains. Results of such a method on translation tasks show that unsupervised alignment methods are not doomed to fail in a cross-domain setting. We obtain BLEU-1 scores ranging from 0.38 to 0.50 on translation tasks, where previous fully unsupervised alignment methods obtain near-zero scores in cross-domain settings.

**Keywords:** Machine learning · Natural Language Processing · Biomedical information · Multilingual representations · Domain adaptation

## 1 Introduction

Word embeddings provide useful representations for many downstream tasks [13] and have been generalized in the cross-lingual context to the concept of Unsupervised Cross-lingual Embedding (UCE) [6, 2]. UCEs learn a distance-preserving transformation, or isometry, mapping one language to the other. This kind of method was shown to work well in some cases, but fails when the embeddings of each language come from a different domain [19]. Our work is motivated by a need for effective UCEs in cross-domain settings. Domain-specific data, such as scientific publications, can be rare in languages other than English. Our goal is to show that UCEs are not doomed to fail in a setting where

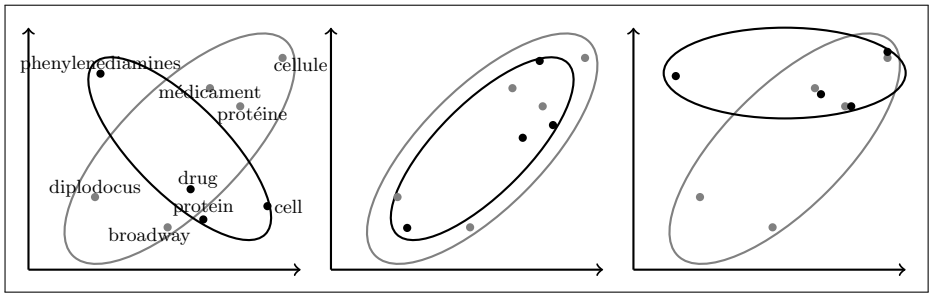



**Fig. 1.** Toy example showing different alignment of a domain (grey) with another (black), the initial unaligned embeddings (left) do not align well when considering all words (center), we aim to align them partially (right)

we have a domain-specific English embedding (e.g. trained on PubMed) and a general-domain non-English embedding (e.g. trained on French Wikipedia).

Methods based on distance-preserving transformations seem to fail in such a cross-domain setting [19]. This suggests that the approximate isometry between embeddings, which is necessary for such methods to work, is not verified in this setting. However, we noticed that alignment methods usually try to align the whole embeddings together, or more precisely a large set of the most frequent words of each embedding, typically 20k [2]. Yet, between two different domains, the distribution of the vocabulary may vary. Some words might be more frequent in one domain and less in the other, or even absent.

We hypothesize that it is possible to improve unsupervised cross-lingual embedding in a cross-domain setting by trying to align well-chosen subsets of each vocabulary. This "partial alignment" of embeddings could be suitable for certain domain-specific tasks. We provide a toy example in Fig. 1 to illustrate this.

Our contribution is threefold: (1) we measure reduced deviation from isometry between parallel vocabularies for embeddings from different domains; (2) we propose an unsupervised alignment method based on a partial alignment outperforming other unsupervised methods in a cross-domain setting; (3) we visualize this partial alignment with t-SNE, a dimensionality-reduction technique [12].

Before detailing our approach in Section 3, we describe the context of our research with isometry-based embedding alignment methods (Section 2.1) and the Bottleneck distance as a measure of the deviation from isometry (Section 2.2). Then we introduce our proposed method for aligning cross-domain embeddings (Section 3), before detailing experiments and results (Section 4).

## 2 Context

### 2.1 Isometry-based Alignment Methods

Shortly after dense word embeddings were introduced [13], it was proposed to map embeddings from distinct languages to a shared space [14]. To perserve the shape of monolingual embeddings, the mapping applied to an embedding must be distance-preserving [14, 6]. Such a transformation is called isometry.

**Definition 1.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two metric spaces. A map $f : \mathcal{X} \to \mathcal{Y}$ is an isometry if for any $(x_i, x_j) \in \mathcal{X}^2$, we have:*

$$d(x_i, x_j) = d(f(x_i), f(x_j)) \tag{1}$$

A linear finite-dimensioned isometry produces an orthogonal matrix, a square matrix whose transpose is its inverse. Supervised methods for cross-lingual embeddings learn an orthogonal mapping between representations of a bilingual dictionary [14]. Following the formalism of [6], we have:

$$W^* = \arg\min_{W \in \mathcal{O}_d} ||AW - B||_F^2 \tag{2}$$

where $A \in \mathbb{R}^{N \times d}$ and $B \in \mathbb{R}^{N \times d}$ are the representations of the $N$ entries of the dictionary in the source and target embeddings of dimension $d$ and $W^*$ is the learned mapping in $\mathcal{O}_d$, the set of orthogonal matrices. Procrustes analysis [18] gives us $W^* = UV^\top$, with the singular value decomposition (SVD) $A^\top B = USV^\top$.

But to learn a mapping in an unsupervised way, given two embeddings $X$ and $Y$, we also need to learn a dictionary as a permutation matrix $P$, which is also an isometry (permutation matrices are orthogonal matrices):

$$W^*, P^* = \arg\min_{P \in \mathcal{P}_n, W \in \mathcal{O}_d} ||XW - PY||_F^2 \tag{3}$$

With the advent of iterative self-learning [1] which alternates between learning the mapping $W$ and the dictionary $P$, alignments required fewer and fewer training pairs of words. It led to fully unsupervised methods with adversarial learning [6] and initialization heuristics. VecMap [2] is one of those self-learning methods which introduces decreasing random noise in the process, inspired by simulated annealing, for more robust alignment. To account for local variations of the density of embedidngs, a Cross-domain Similarity Local Scaling (CSLS) criterion [6] is often used [6, 2, 8], leveraging the average distance $d$ of $k$ nearest neighbors[1]:

$$CSLS(u, v) = d(u, v) - \frac{1}{k} \sum_{x \in \mathcal{N}_k(u)} d(x, u) - \frac{1}{k} \sum_{y \in \mathcal{N}_k(v)} d(y, v) \tag{4}$$

But for $X$ and $Y$ to align correctly with such methods, they must be approximately isometric. Unfortunately, it was shown that UCEs relying on an

[1] usually $k = 10$

orthogonal mapping need three conditions to give accurate results [19]: (1) languages must be morphologically similar; (2) the monolingual training corpora must be from the same domain; and (3) the same model must be used (CBOW Spanish embeddings did not align with Skip-gram English). These drawbacks eventually led to several methods featuring a weak orthogonality constraint [16]. But we hypothesize that, when we have embeddings from different domains, the isometry assumption must not be completely abandonned, as there might still be approximate isometry between relevant subsets of those embeddings.

### 2.2 Measuring Deviation from Isometry

To verify our hypothesis, we need a way to measure the deviation from isometry of two metric spaces. First, we must be able to evaluate to what extent two aligned sets coincide. For that we can rely on the Hausdorff distance.

**Definition 2.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two compact subsets of a metric space $(\mathcal{Z}, d_Z)$. The Hausdorff distance is defined by:*

$$d_H^{\mathcal{Z}}(\mathcal{X}, \mathcal{Y}) = \max\left(\sup_{x\in\mathcal{X}} \inf_{y\in\mathcal{Y}} d_Z(x,y), \sup_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} d_Z(x,y)\right) \tag{5}$$

Intuitively, the Hausdorff distance is the maximum distance between pairs of nearest neighbors. From that we can build the Gromov-Hausdorff distance, which gives a theoretical measure of the deviation from isometry.

**Definition 3.** *Let $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$ be two metric spaces. The Gromov-Hausdorff distance is defined by:*

$$d_{GH}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) = \inf_{\mathcal{Z}, f, g} d_H^{\mathcal{Z}}(f(\mathcal{X}), g(\mathcal{Y})) \tag{6}$$

*With $f : \mathcal{X} \to \mathcal{Z}$ and $g : \mathcal{Y} \to \mathcal{Z}$ isometries matching both metric spaces to a single metric space $(\mathcal{Z}, d_Z)$.*

The Gromov-Hausdorff distance is the minimum over all isometric transformations of the Hausdorff distance. It measures how well two metric spaces can be aligned without deforming them. This distance, intractable to compute in practice, needs to be approximated. Several works use diverse metrics. A similarity based on a spectral analysis of neighborhood graphs is correlated with performances of alignment methods [19]. Earth Mover's Distance between embeddings was linked with typological similarity between languages [20], showing why distant language pairs are more difficult to align. Another metric is the Bottleneck distance between the persistence diagrams of the Rips complex filtrations of each metric space and it was shown to be a tight lower-bound for the Gromov-Hausdorff distance [5], and was found to better correlate with the ability to align embedding with an orthogonal mapping than previously mentioned metrics [16].

For a more formal definition of persistence diagram, Rips complex and filtrations, the reader might refer to relevant literature [5]. In our case, we create a

parameter $t$ that varies from 0 to $+\infty$, and compute a graph for each embedding such that two points of an embedding that are at a distance smaller than $2t$ are linked by an edge. This graph is a Rips complex, or rather a simplified version of it because we only look at connected components, hence only edges, not higher-dimensional simplexes. We start with as many connected components as elements in the embedding ($t = 0$) and gradually decrease their number by merging them. This sequence of Rips complexes is called a filtration, on which we can compute a persistence diagram for each embedding: a list of points $(t_{\text{birth}}, t_{\text{death}})$ for each connected component that appears during the filtration recording the $t{=}t_{\text{birth}}$ at which it appears and the $t{=}t_{\text{death}}$ at which it is merged with another.

Comparing the persistence diagrams for two embeddings allows us to measure to what extent they differ topologically. This is the Bottleneck distance.

**Definition 4.** *Given two multi-sets $A$ and $B$ in $\overline{\mathbb{R}}^2$, the Bottleneck distance is defined by:*

$$d_B^{\infty}(A, B) = \min_{\gamma} \max_{p \in A} ||p - \gamma(p)||_{\infty} \tag{7}$$

*With $\gamma$ ranging over bijections between $A$ and $B$.*

The Bottleneck distance between Rips filtration of metric spaces gives us a lower bound for the Gromov-Hausdorff distance.

**Theorem 5.** *From [5]. For any finite metric spaces $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$ and for any $k \in \mathbb{N}$:*

$$d_B^{\infty}(\mathcal{D}_k\mathcal{R}(\mathcal{X}, d_X), \mathcal{D}_k\mathcal{R}(\mathcal{Y}, d_Y)) \leq d_{GH}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) \tag{8}$$

To make it simpler, in the following, when we mention the "Bottleneck distance" between two sets of embeddings, we will be actually referring to the Bottleneck distance between the persistence diagrams of the filtrations of Rips complexes built over those embeddings. We will use this Bottleneck distance in our first experiment to measure the deviation from isometry between diverse subsets of embeddings from different languages and domains, showing that some subsets of embeddings from different domains are more topologically similar than the whole embeddings themselves.

## 3 Proposed Approach

The proposed hypothesis states that embeddings from different domains could still be partially aligned. So the biggest challenge is to find the relevant subsets that align well. We propose here a method based on a proxy embedding of the same domain as one embedding and the same language as the other.

The proposed approach can be summarized as follows: we align the source embedding and the proxy embedding, which are from the same domain, with an isometry-based alignment method. We build a dictionary from it that takes into account the target vocabulary and use this dictionary to find a mapping

between the source and target. The alignment between source and proxy should work since both embeddings are from same domain and the filtered dictionary used to align the source and the target should allow this partial alignment and avoid aligning subsets of vocabulary that are not relevant to each other.

More formally, as shown in Algorithm 1, we have $X$ and $Z$ two embeddings of distinct languages and domains (e.g. French Wikipedia and English PubMed). Let $Y$ be a proxy embedding of same domain as $X$ and same language as $Z$ (English Wikipedia in our example). The proposed method aligns $X$ and $Y$ together (of same domain) by solving Equation 3, giving aligned embeddings $\widetilde{X}$ and $\widetilde{Y}$. We use the VecMap algorithm [2], a state-of-the-art unsupervised method based on self-learning as in Section 2.1, which performs well on same-domain embeddings. We then compute a dictionary by nearest-neighbor search between elements of $\widetilde{X}$ and $\widetilde{Y}$ using the CSLS criterion. We filter the dictionary by keeping only the pairs for which the entry word for the target language is in the vocabulary of $Z$. And finally, we align $X$ and $Z$ by learning the orthogonal mapping matching the embeddings of the entries $A$ and $B$ of the filtered dictionary (Equation 2). Using Procrustes analysis, it can be obtained as $W^* = UV^\top$ with the result of the SVD $A^\top B = USV^\top$.

We have $XW^*$ and $Z$ the aligned cross-domain and cross-lingual embeddings.

This cross-domain and cross-lingual alignment method could serve as initialization for unsupervised translation models [11, 3] and we will evaluate it on two domain-specific translation tasks in the following section.

**Algorithm 1:** Proposed method of alignment with proxy

**Input** : source embedding $X$, target $Z$ and proxy $Y$.
**Output** : aligned source and target $XW^*$ and $Z$.
Create aligned $\widetilde{X}$ and $\widetilde{Y}$ with VecMap [2] on $X$ and $Y$.
Initialize $D = \{\}$
**for** *word* $w$ *in* $\widetilde{X}$ **do**
  Find $w'$ nearest-neighbor of $w$ in $\widetilde{Y}$.
  **if** $w'$ *in vocabulary of* $Z$ **then**
    Add $(w, w')$ to dictionary $D$.
  **end**
**end**
**for** *word* $w'$ *in* $\widetilde{Y}$ *and vocabulary of* $Z$ **do**
  Find $w$ nearest-neighbor of $w'$ in $\widetilde{X}$.
  Add $(w, w')$ to dictionary $D$.
**end**
Build $A \in \mathbb{R}^{N \times d}$ and $B \in \mathbb{R}^{N \times d}$ embeddings entries of $D$ in $X$ and $Z$.
Perform SVD: $A^\top B = USV^\top$.
Solve Equation 2: $W^* = UV^\top$.
Compute aligned embeddings $XW^*$ and $Z$.

## 4 Experimental Validation

To assess our hypothesis we perform three experiments: (1) showing that some vocabulary subsets of embeddings from different domains are topologically similar with Bottleneck distance; (2) evaluating the performance of the proposed partial alignment method on a translation task; (3) visualizing the learned partial alignment with a dimensionality reduction technique.

### 4.1 Datasets

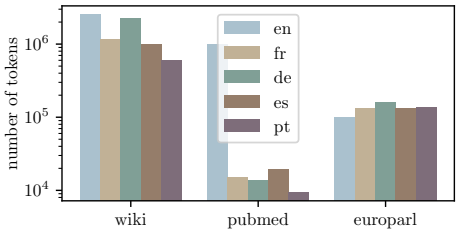



**Fig. 2.** Number of distinct tokens in each embedding by corpus and by language

For all experiments, we leverage embeddings built on three different corpora (Wikipedia, PubMed, EuroParl) with five different languages (English, French, German, Spanish and Portuguese). PubMed is a collection of approximately 21 million biomedical abstracts[2] mainly written in English. EuroParl [10] is a parallel corpus built from proceedings of the European Parliament. All embeddings were built with FastText[4] with 300 dimensions. Those from Wikipedia were obtained directly from the FastText website and we trained FastText ourselves on PubMed and EuroParl using the official implementation. We show on Fig. 2 the vocabulary size for each embedding. Wikipedia embeddings as well as the English PubMed embeddings have a vocabulary size of the same order of magnitude. The other corpora are not comparable. This allows us to emphasize on why we need cross-domain alignment methods as domain-specific data is sometimes lacking in languages other than English.

### 4.2 Measuring the Deviation from Isometry

We use the Bottleneck distance (Section 2.2) to measure the deviation from isometry between various subsets of different pairs of embeddings. We limit the subsets to 5k words for computability reasons and because previous works [16] have used the same constant and shown correlation with the ability to align embeddings. We compute the Bottleneck distance between three kinds of subsets:

[2] Courtesy of the U.S. National Library of Medicine https://www.nlm.nih.gov/databases/download/pubmed_medline.html

**Table 1.** Bottleneck distance between the 5k most frequent words for different subsets and different domain pairs averaged over all language pairs with English (standard deviation between parenthesis)

| Most frequent words of each embedding | | English | | |
|---|---|---|---|---|
| | | wiki (1) | Pubmed (2) | Europarl (3) |
| fr/de/es/pt | wiki (a) | 0.08 (0.03) | 0.12 (0.02) | 0.15 (0.05) |
| | Pubmed (b) | 0.21 (0.08) | 0.17 (0.06) | 0.23 (0.08) |
| | Europarl (c) | 0.19 (0.03) | 0.16 (0.01) | 0.05 (0.02) |
| Most frequent words in PubMed parallel corpus | | English | | |
| | | wiki | Pubmed | Europarl |
| fr/de/es/pt | wiki (d) | 0.08 (0.02) | 0.08 (0.01) | 0.07 (0.03) |
| | Pubmed (e) | 0.20 (0.06) | 0.16 (0.07) | 0.23 (0.08) |
| | Europarl (f) | 0.14 (0.03) | 0.10 (0.03) | 0.07 (0.01) |
| Most frequent words in Europarl parallel corpus | | English | | |
| | | wiki | Pubmed | Europarl |
| fr/de/es/pt | wiki (g) | 0.09 (0.04) | 0.10 (0.03) | 0.13 (0.04) |
| | Pubmed (h) | 0.18 (0.06) | 0.16 (0.07) | 0.21 (0.06) |
| | Europarl (i) | 0.17 (0.02) | 0.16 (0.03) | 0.05 (0.02) |

- The 5k most frequent words for each embedding, which was shown to correlate with the ability to align whole embeddings [16];
- The 5k most frequent words in each language of a parallel Pubmed corpus;
- The 5k most frequent words in each language of a parallel Europarl corpus.

We measure the Bottleneck distance for all domain pairs for language pairs involving English only, as we were able to obtain only parallel copora between English and other languages. Results are reported on Table 1. We average over the other languages to summarize, as results were consistent across languages as shown by the low standard deviations, except for pairs including non-English Pubmed embeddings for which the small size of those embeddings (Fig. 2) might explain the high Bottleneck distance and standard deviation. The lines corresponding to the latter situation (b,e,h) are discarded from further analysis.

The Bottleneck distance for same-domain pairs and most frequent words for Wikipedia and Europarl (a1,c3) indicates that topologically similar subsets have a Bottleneck distance below 0.09 whereas disimilar ones, such as cross-domain pairs have Bottleneck distances above 0.10. This verifies that embedding as a whole cannot be aligned if not from the same domain.

For each cross-domain pairs, the Bottleneck distance for parallel subsets is lower than for frequent words or at least equal. The most striking cases are when comparing non-English Wikipedia with English Europarl or Pubmed on the Pubmed parallel vocabulary (d2,d3), where the Bottleneck distance becomes comparable to those of same-domain pairs. For other domain-pairs, the Bottleneck distance is higher but we can still observe decreases with respect to the first subset (comparing c1 with f1,i1 and c2 with f2,i2). This suggests that the embeddings of parallel vocabularies for cross-domain pairs are topologically close.

With this first experiment, we can support our hypothesis that the quasi-isometry assumption might still hold between well-chosen subsets of embeddings from different domains. Indeed, the Bottleneck distance for cross-domain pairs was systematically smaller for parallel vocabularies. However, in an unsupervised method, such subsets may not be straightforward to find. That is why we devised the method described in Section 3, which we evaluate in the following section.

### 4.3 Unsupervised Cross-Domain and Cross-Lingual Alignement

The method proposed in Section 3 is evaluated on the Europarl parallel corpora and on the test set of the Biomedical WMT19 dataset[3]. This latter dataset consists of non-English Pubmed abstracts (around 100 by language pair) and their English translation. We proceed token-by-token by nearest-neighbor search with cosine distance and CSLS criterion. We use the BLEU-1 score [15] to evaluate translations:

$$\text{BLEU-1}(\mathrm{r},\mathrm{h}) = \min\left(1, e^{1-\frac{|\mathrm{r}|}{|\mathrm{h}|}}\right) \frac{\sum_{w\in\mathrm{h}} \min(\mathrm{count}_\mathrm{h}(w), \mathrm{count}_\mathrm{r}(w))}{\sum_{w\in\mathrm{h}} \mathrm{count}_\mathrm{h}(w)} \tag{9}$$

It is a modified precision measure on the tokens with an additional term penalizing short translations and a clipping on the count of occurences to avoid giving high scores to candidate translations (h) which repeat words from the reference translation (r).

We choose translation over bilingual lexicon induction, which does not account for morphological variations of words [7] and gives too much importance to specific words such as proper nouns [9]. Moreover, cross-lingual embeddings can be used as initialization models in unsupervised translation models [11, 3].

**Table 2.** BLEU-1 scores on Biomedical WMT19 and Europarl

| cross-domain | Biomedical WMT19 | | | | Europarl | | | |
|---|---|---|---|---|---|---|---|---|
| | fr-en | es-en | de-en | pt-en | fr-en | es-en | de-en | pt-en |
| MUSE | 0.05 | 0.06 | 0.06 | 0.07 | 0.00 | 0.03 | 0.01 | 0.01 |
| WP | 0.08 | 0.08 | 0.05 | 0.05 | 0.01 | 0.01 | 0.01 | 0.01 |
| VecMap (unsupervised) | 0.09 | 0.06 | 0.07 | 0.07 | 0.02 | 0.02 | 0.03 | 0.02 |
| VecMap (weakly supervised) | 0.30 | 0.37 | 0.25 | 0.28 | 0.33 | 0.39 | 0.31 | 0.33 |
| proposed method | **0.38** | **0.50** | **0.31** | **0.47** | **0.40** | **0.44** | **0.37** | **0.44** |
| same-domain | Biomedical WMT19 | | | | Europarl | | | |
| | fr-en | es-en | de-en | pt-en | fr-en | es-en | de-en | pt-en |
| MUSE | 0.43 | 0.58 | **0.40** | 0.53 | 0.43 | 0.47 | **0.41** | 0.43 |
| WP | 0.45 | 0.57 | 0.36 | 0.51 | 0.45 | 0.47 | 0.40 | 0.46 |
| VecMap (unsupervised) | **0.46** | **0.58** | 0.37 | **0.56** | **0.46** | **0.49** | 0.41 | **0.47** |
| *UCAM* (supervised) | - | *0.71* | *0.61* | - | - | - | - | - |

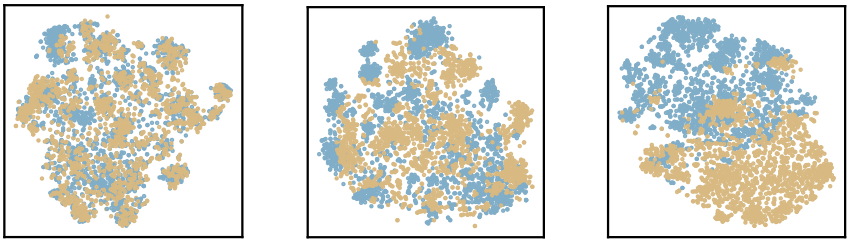
[3] http://www.statmt.org/wmt19/biomedical-translation-task.html

Results for translation on the Biomedical WM19 and Europarl tasks are shown on Table 2. The proposed method is compared to the unsupervised methods VecMap[2], MUSE[6] and Wasserstein-Procrustes (WP) [8] applied to cross-domain embeddings, a weakly supervised method based on identical words and VecMap applied to cross-domain embeddings as well, the same unsupervised methods (VecMap, MUSE, WP) applied to same-domain embeddings, and the UCAM submission [17] to the Biomedical WMT19 challenge, a supervised deep learning model which gives an upper-bound baseline.

In all languages and domains, the proposed method outperforms any other that tries to align embeddings from different domains. In such cross-domain settings, fully unsupervised methods obtain near-zero translation scores, whereas the proposed unsupervised method gives fair results, even outperforming the weakly-supervised method based on identical words. The proposed method demonstrates that it is possible to align embeddings from different domains in an unsupervised manner, although it is outperformed by same-domain alignments.

Nevertheless, the fact that we were able to obtain fair translation scores with a strictly isometric mapping suggests that our hypothesis of isometry between subsets might hold. To further validate this hypothesis and the intuition behind our method, we show in the following section with a visualization method that our alignment is indeed partial.

### 4.4 Visualization with t-SNE



(a) VecMap, same domain. (b) VecMap, cross-domain. (c) ours (cross-domain).

**Fig. 3.** t-SNE for alignment between French Wikipedia and English Wikipedia (same domain) or English Pubmed (cross-domain)

We use a dimensionality-reduction technique, t-SNE [12], to confirm that the proposed method performs a partial alignment. Results for the French-English pair are shown on Fig. 3. Blue points are t-SNE representations of embeddings of French words from Wikipedia embedding and orange points are for English words either from Wikipedia (3(a)) or Pubmed (3(b),3(c)) embedding, represented after being aligned by VecMap (3(a),3(b)) or the proposed method (3(c)). A classic alignment method like VecMap aligns the embedding as a whole. In the

same-domain setting (3(a)) it works well as suggested by our previous results in the translation tasks and the seemingly local alignment of small clusters. In the cross-domain setting (3(b)), we do not observe the same local alignment, which is corroborated by the near-zero BLEU-1 score in the previous experiment. Finally, t-SNE for the proposed method (3(c)) seems to show that there is a partial alignment. And even if we do not observe local alignment of clusters as in VecMap for same domain, we can confirm with this visualization that our filtered dictionary allows a partial alignment as hypothesized in introduction (Fig. 1).

## 5 Discussion and Conclusion

We proposed a novel method for aligning embeddings from different domains, a setting where other unsupervised methods failed. By aligning only a well-chosen subsets instead of the whole embeddings, we showed that unsupervised isometry-based alignment were not doomed to fail in such setting. However, our method needs additional data. Further work could try to improve on the proposed method by removing this need for additional data or to use it as initialization of an unsupervised neural machine translation model as is already done with same-domain alignments [11, 3]. Another limitation of our work is that the Bottleneck distance is only a lower bound of the Gromov-Hausdorff distance. Although this bound is expected to be tight [5], to the best of our knowledge, there is no formal proof of this tightness. Nevertheless, there is empirical evidence of correlation between Bottleneck distance and the ability to align embeddings with an orthogonal mapping [16], although this was only shown for different pairs of languages in the same domain. Further work would try to demonstrate the same correlation for different domains.

We showed with Bottleneck distance that some subsets of embeddings were topologically similar, despite being from different languages and domains. We demonstrated that orthogonal mapping alignments are not doomed to fail thanks to the proposed method which outperforms any other unsupervised distance-preserving alignment applied to embeddings from different domains. Finally, we confirmed that the proposed method was performing the hypothesized partial alignment with a dimensionality-reduction technique. With these three pieces of evidence, we are confident that the approximate isometry assumption can still hold between well-chosen subsets of cross-domain embeddings.

## References

1. Artetxe, M., Labaka, G., Agirre, E.: Learning bilingual word embeddings with (almost) no bilingual data. In: Proceedings of the 55th Annual Meeting of ACL (2017). https://doi.org/10.18653/v1/P17-1042
2. Artetxe, M., Labaka, G., Agirre, E.: A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In: Proceedings of the 56th Annual Meeting of ACL (2018). https://doi.org/10.18653/v1/P18-1073

3. Artetxe, M., Labaka, G., Agirre, E.: An effective approach to unsupervised machine translation. In: Proceedings of the 57th Annual Meeting of ACL (2019). https://doi.org/10.18653/v1/P19-1019
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of ACL pp. 135–146 (2017), https://www.aclweb.org/anthology/Q17-1010
5. Chazal, F., Cohen-Steiner, D., Guibas, L.J., Mémoli, F., Oudot, S.Y.: Gromov-hausdorff stable signatures for shapes using persistence. In: Proceedings of the Symposium on Geometry Processing. p. 1393–1403. SGP '09, Eurographics Association, Goslar, DEU (2009)
6. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data (2017), http://arxiv.org/abs/1710.04087
7. Czarnowska, P., Ruder, S., Grave, E., Cotterell, R., Copestake, A.: Don't forget the long tail! a comprehensive analysis of morphological generalization in bilingual lexicon induction. In: Proceedings of EMNLP-IJCNLP (2019). https://doi.org/10.18653/v1/D19-1090
8. Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., Grave, E.: Loss in translation: Learning bilingual word mapping with a retrieval criterion. In: Proceedings of the 2018 Conference of EMNLP (2018). https://doi.org/10.18653/v1/D18-1330
9. Kementchedjhieva, Y., Hartmann, M., Søgaard, A.: Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In: Proceedings of EMNLP-IJCNLP (2019). https://doi.org/10.18653/v1/D19-1328
10. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Conference Proceedings: the tenth Machine Translation Summit. AAMT, AAMT, Phuket, Thailand (2005), http://mt-archive.info/MTS-2005-Koehn.pdf
11. Lample, G., Conneau, A., Denoyer, L., Ranzato, M.: Unsupervised machine translation using monolingual corpora only. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=rkYTTf-AZ
12. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013), http://arxiv.org/abs/1301.3781
14. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation (2013), http://arxiv.org/abs/1309.4168
15. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of ACL (2002), https://doi.org/10.3115/1073083.1073135
16. Patra, B., Moniz, J.R.A., Garg, S., Gormley, M.R., Neubig, G.: Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In: Proceedings of the 57th Annual Meeting of ACL (2019). https://doi.org/10.18653/v1/P19-1018
17. Saunders, D., Stahlberg, F., Byrne, B.: UCAM Biomedical Translation at WMT19: Transfer Learning Multi-domain Ensembles. In: Proceedings of the Fourth Conference on Machine Translation (2019), http://www.aclweb.org/anthology/W19-5421
18. Schönemann, P.H.: A generalized solution of the orthogonal procrustes problem. Psychometrika **31**(1), 1–10 (Mar 1966), https://doi.org/10.1007/BF02289451
19. Søgaard, A., Ruder, S., Vulić, I.: On the limitations of unsupervised bilingual dictionary induction. In: Proceedings of the 56th Annual Meeting of ACL (2018). https://doi.org/10.18653/v1/P18-1072
20. Zhang, M., Liu, Y., Luan, H., Sun, M.: Earth mover's distance minimization for unsupervised bilingual lexicon induction. In: Proceedings of the 2017 Conference on EMNLP (2017). https://doi.org/10.18653/v1/D17-1207