Topic-Based Unsupervised and Supervised Dictionary Induction

YUZHI LIU and MASSIMO PICCARDI, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

Word translation is a natural language processing task that provides translation between the words of a source and a target language. As a task, it reduces to the induction of a bilingual dictionary, which is typically performed by aligning word embeddings of the source language to word embeddings of the target language. To date, all the existing approaches have focused on performing a single, global alignment in word embedding space. However, semantic differences between the various languages, in addition to differences in the content of the corpora used for training the word embeddings, can hinder the effectiveness of such a global alignment. For this reason, in this article we propose conducting the alignment between the source and target embedding spaces by multiple mappings at topic level. The experimental results show that our approach has been able to achieve an average accuracy improvement of +3.30 percentage points over a state-of-the-art approach in unsupervised dictionary induction from languages as diverse as German, French, Italian, Spanish, Finnish, Turkish, and Chinese to English, and +3.95 points average improvement in supervised dictionary induction.

CCS Concepts: • **Computing methodologies** → **Machine translation**; • **Information systems** → **Document topic models;**

Additional Key Words and Phrases: Word translation, dictionary induction, topic-based dictionary induction, topic modeling, word embedding alignment

ACM Reference format:
Yuzhi Liu and Massimo Piccardi. 2023. Topic-Based Unsupervised and Supervised Dictionary Induction. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 22, 3, Article 77 (March 2023), 21 pages.
https://doi.org/10.1145/3564698

1 INTRODUCTION

Word translation is a natural language processing task that provides translation between the words of two given languages. While less general than conventional machine translation, it can be useful to provide first-cut translation of sentences in the absence of a proper translation model, or even for data augmentation of small parallel training corpora. As a task, word translation is equivalent to the induction of a bilingual dictionary and has attracted substantial, recent research [3, 7, 13, 15, 18, 21]. De facto, all approaches proposed to date have framed the induction of the bilingual dictionary as a problem of word embedding alignment between the word embeddings of the source and target language.

Authors' address: Y. Liu and M. Piccardi, Faculty of Engineering and Information Technology, University of Technology Sydney, P.O. Box 123, Broadway, NSW 2007, Australia; emails: lyz15972107087@gmail.com, Massimo.Piccardi@uts.edu.au.

2375-4699/2023/03-ART77 $15.00
https://doi.org/10.1145/3564698

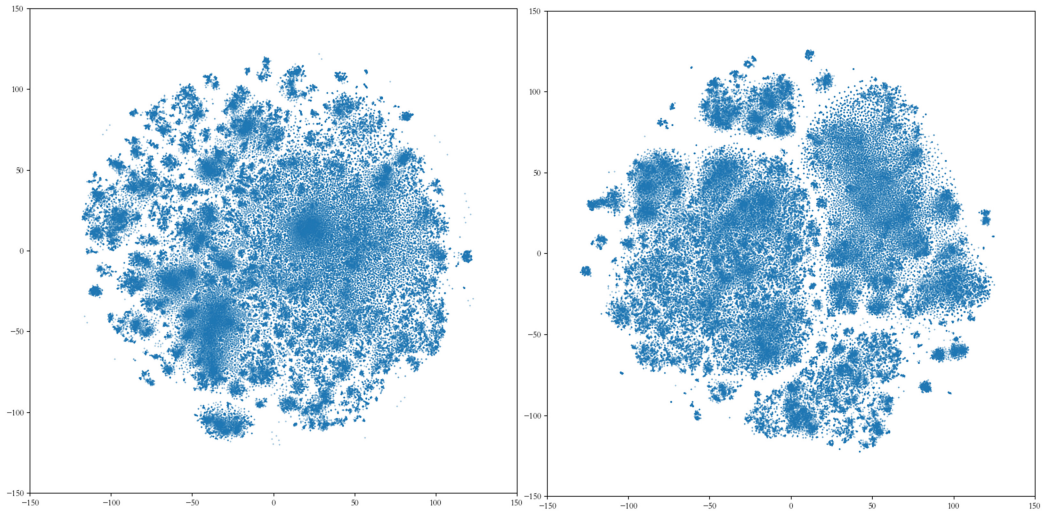

Fig. 1. Word embeddings from an English (left) and a Chinese (right) vocabulary. The embeddings have been projected to 2D using t-SNE [26].

The importance of word embeddings for **natural language processing** (**NLP**) simply cannot be overstated. Popular, non-contextual word embeddings such as word2vec [22], GloVe [23], and fastText [5] have found ubiquitous application in NLP tasks. Notably, Mikolov et al. [21] also observed that the word embedding spaces of different languages can be put in an approximate linear relation. Specifically, they noted that the relative spatial positions of word embeddings are similar across languages; therefore, it is possible to identify a single, linear transformation to map the word embedding space of a language to another's. Once aligned, the two word embedding spaces can be used to carry out nearest-neighbor searches and induce the bilingual dictionary. Based on this idea, a number of unsupervised and "supervised" (i.e., leveraging a small seed dictionary) approaches have been proposed in the literature. However, with typical accuracies ranging between 40 and 80 percentage points, there seems to still be significant room for improvement.

While the notion of a linear, global alignment between the word embedding spaces is certainly suggestive, it may fail to accurately model the mapping due to semantic differences between the languages and also the finiteness of the corpora used for training the word embeddings. To give an idea of the differences between the word embedding spaces of different languages, Figure 1 shows a plot for corresponding English and Chinese vocabularies (projected to 2D with t-SNE [26]). The distribution of the English embeddings appears relatively uniform, while the Chinese embeddings are split into approximately four "macroregions" by evident gaps. For this reason, in this article we propose a novel dictionary induction approach that leverages *a set* of linear mappings. Differently from existing techniques, the proposed approach uses a partition of the source vocabulary created using a topic model, and performs the mapping at topic level. In a range of experiments on unsupervised and supervised dictionary induction, the proposed approach has achieved a marked average accuracy improvement over a state-of-the-art approach [7].

The remainder of this article is organized as follows: Section 2 reviews the related work on word translation and dictionary induction. Section 3 describes the main stages of the dictionary induction task, while Section 4 introduces the proposed approach. Section 5 presents an ample range of experiments with a detailed analysis of the results. Finally, Section 6 concludes the article.

2 RELATED WORK

The field of word translation by word embedding alignment was initiated by Mikolov et al. [21], stemming from the remarkable observation that different languages exhibit comparable word embedding spaces. Mikolov et al. [21] also proposed determining the mapping matrix by minimizing the average L2 distance between the mapped source embeddings and the target embeddings of their translations. This approach requires complete supervision of the correspondence between words in the two languages. Later, Xing et al. [30] proposed two improvements: the normalization of the word embeddings, and an orthogonality constraint on the mapping matrix. Constraining the mapping matrix to be orthogonal makes the alignment problem a Procrustes problem [24], that can be solved by simply applying singular value decomposition. Since both these approaches require abundant parallel data at word level, later methods have focused on reducing the annotation effort. For instance, Vulić and Korhonen [27] have proposed a model called HYBWE that uses an inexpensive seed bilingual dictionary, extracted from documents aligned using an auxiliary model [29]; Smith et al. [25] have proposed building a seed training dictionary by leveraging identical character strings in the two languages (an approach that can only be applied to languages sharing a common alphabet); and Artetxe et al. [2] have proposed VecMap, an iterative method that can use either as little as 25 word pairs or a generated list of numerals to supervise the training.

To the best of our knowledge, the approach presented in [20] has been the first to achieve word embedding alignment without any parallel data. The author trained the mapping matrix with a **generative adversarial network** (**GAN**) [12] aimed to discriminate between the mapped source embeddings and the target embeddings. Similarly, Zhang et al. [33] also performed the alignment of the word embedding spaces using a GAN, showing improvements over [20]. However, neither model has reported accuracies comparable to those of supervised approaches. In fact, the first unsupervised model that has achieved accuracies in the range of those of supervised models on some language pairs has been MUSE [7]. MUSE first uses a GAN to roughly align the source word embedding space to the target space. Then, it selects matching pairs to produce a seed dictionary, which it uses to refine the mapping matrix with Procrustes. The selection and refinement are iterated a number of times, until a validation criterion is met. In addition, MUSE has proposed and leveraged a novel pairwise distance between word embeddings, called **cross-domain similarity local scaling** (**CSLS**), to mollify the reported "hubness" of word embedding spaces [10].

A few unsupervised approaches that do not rely on GANs for initialization have also been proposed. Artetxe et al. [3] have proposed an updated version of VecMap that uses a weak initialization of the mapping, and fine-tuning via stochastic dictionary induction and symmetric re-weighting. Vulić et al. [28] have proposed an improvement to VecMap that generalizes the notion of first- and second-order similarity of the monolingual embeddings to *n*-th-order similarity. Hoshen and Wolf [15] have derived the mapping matrix with a PCA-based initialization and a **mini-batch cycle iterative closest point** (**MBC-ICP**) algorithm. In turn, Grave et al. [13] have proposed an unsupervised approach that finds the correspondences between source and target embeddings by minimizing their Wasserstein distance. However, according to Hartmann et al. [14], none of these models has achieved accuracies comparable to those of MUSE augmented with stochastic dictionary induction. More recently, Li et al. [18] have improved the robustness of VecMap by reducing the dimensionality of the word embeddings with PCA, and Cao and Zhao [6] have replaced the dimensionality reduction step with affine transforms (rotation and scaling), both reporting improvements over the original model.

As part of the related literature, JP et al. [16] have investigated the possibility to even *predict* the word embeddings of a low-resource target language based on the embeddings of a high-resource source language and a mapping function (rather than learning them from monolingual text as

usual). In addition, other studies have investigated the usefulness of topic information for machine translation. For instance, Xiong and Zhang [32] have proposed a translation model that explicitly extracts the topics from each source sentence to use them to guide the translation. At their turn, Xiong et al. [31] have proposed three term translation models that use topic information to improve disambiguation, consistency, and "unithood" in machine translation. While other work in the area exists, we are not aware of any previous studies that have focused on topic-based word translation.

3 BACKGROUND

Existing methods for dictionary induction mainly focus on aligning the two word embedding spaces globally. The task breaks down into two fundamental stages: (1) the inference of a mapping matrix, W, that is used to map the words from the source space to the target space, and (2) the matching of the mapped source words to the words of the target language (word matching, for short). These two stages may be iterated multiple times until a validation criterion is met. If a small set of manually aligned word pairs from the two languages is provided, the approach is classified as supervised; otherwise, it is classified as unsupervised. In the following, we briefly review the two main stages.

3.1 Inference of the Mapping Matrix

Given a set of words in the source embedding space, X, and a set of matching words from the target embedding space, Y (essentially, a supervised parallel dictionary), the goal of the mapping stage is to learn a mapping matrix, W, minimizing the L2 distance between the mapped source embeddings, WX, and the target embeddings, Y [7]:

$$W^* = \underset{W}{\operatorname{argmin}} \|WX - Y\|. \tag{1}$$

The solution of (1) can be found with any standard least-square solver. However, Xing et al. [30] have showed that constraining the mapping matrix, W, to be an orthogonal matrix ($W \in \mathrm{O}$) can generally improve the word translation accuracy. Therefore, (1) is converted into a Procrustes problem, which can be solved by simply performing the **singular value decomposition** (**SVD**) of matrix product YX^T:

$$\begin{aligned} W^* &= \underset{W \in \mathrm{O}}{\operatorname{argmin}} \|WX - Y\| \\ &= UV^T, \text{ where } U\Sigma V^T = \mathrm{SVD}(YX^T). \end{aligned} \tag{2}$$

The inference of the mapping matrix requires an aligned bilingual dictionary. In the supervised case, a small aligned dictionary (the "seed" dictionary) is manually provided. In the unsupervised case, the dictionary can be constructed by aligning the distributions of the source and word embeddings using, for instance, GANs [7], low-dimensional **iterative closest point** (**ICP**) optimization [15], Gromov-Wasserstein distance minimization [1, 3], or the Gold-Rangarajan relaxation [13].

3.2 Word Matching

Once the mapping matrix has been determined, it can be used to map any source word embedding, x, to the target space, and search for a suitable matching word, y. Therefore, this stage can be framed as a nearest-neighbor search using the L2 distance or cosine similarity between mapped source embeddings and target embeddings. However, Dinu and Baroni [10] and many other works have remarked that word embedding spaces suffer from the "hubness" problem, i.e., the fact that some words have many close neighbors while others have none, affecting the word matching. To mollify this problem, Conneau et al. [7] have proposed using a normalized similarity measure,

called CSLS. Given two word embeddings, x and y, from the source and target languages, respectively, CSLS can be expressed as

$$\begin{aligned} &\text{CSLS}(Wx, y) \\ &= 2\cos(Wx, y) - r_T(Wx) - r_S(y), \end{aligned} \tag{3}$$

where $\cos(Wx, y)$ is the cosine similarity between a mapped source embedding, Wx, and a target embedding, y; $r_T(Wx)$ is the average cosine similarity between Wx and its k-nearest target embeddings; and $r_S(y)$ is the average cosine similarity between y and its k-nearest mapped source embeddings. Conneau et al. [7] have shown that CSLS is able to compensate for the "hubness" problem and used it to find matching pairs. In addition, they have imposed that the paired words be mutual neighbors (y has to be the closest neighbor of Wx, and vice versa) and have a minimum frequency in their respective sets. Since this approach has been used for word matching in both [7, 13], we have adopted it also for our work.

4 PROPOSED APPROACH

The idea of a single, global mapping between the source and target word embeddings seems well justified by the general properties of word embedding spaces [21]. However, semantic differences between languages, as well as differences in the corpora used to train the embeddings, may cause structural differences to their word embedding spaces which may not be properly accounted for by a simple linear mapping. To illustrate this, in Figure 2 we show the embeddings of several words from four different topics (*animals*, *colors*, *education*, and *months*) in English and Italian after being aligned linearly using MUSE [7]. It is clear that words from the two languages have similar geometric arrangements at global level. However, when the individual topics are inspected in greater detail, the differences become noticeable. For instance, in the *animals* topic the Italian word "cervo" ("deer" in English) is closer to the English word "sheep" rather than "deer." These local differences in word embedding spaces can significantly affect the induction of the bilingual dictionary. For this reason, we propose aligning the word embeddings at topic level, relying on the intuition that a linear mapping may be more suited to map words within the same topic. Figure 3 shows an architecture diagram of the proposed approach: in the first step, the source word embeddings are clustered by topic (Section 4.1), and in the second step a mapping matrix is inferred separately for each cluster (Section 4.2). We present these two steps in detail in the remainder of this section.

4.1 Clustering of the Source Embeddings

The word embeddings of the source language can be clustered by leveraging topic models that operate in embedding spaces. Notably, these include Gaussian LDA [8] and the **embedded topic model** (**ETM**) [9]. However, even conventional statistical clustering approaches such as the **Gaussian mixture model** (**GMM**) [19] can be used to cluster the word embeddings. For our experiments, we have adopted Gaussian LDA and the GMM, and we thus briefly recap them in the following.

4.1.1 Gaussian LDA. **Latent Dirichlet Allocation** (**LDA**) is one of the most popular models for inferring the topic structure of document corpora [4]. LDA models both the topics and the words in the documents as categorical variables, and assumes Dirichlet priors over their distributions. Conversely, Gaussian LDA [8] models the words as numerical embeddings and assumes them normally distributed within each topic. In turn, the parameters of the normal distributions of each topic are treated as random variables, with a learnable normal prior over the mean ($N(\mu_k|\mu, \Sigma_k)$) and a learnable Inverse Wishart prior over the covariance ($W^{-1}(\Sigma_k|\Psi, \nu)$). Algorithm 1 shows the generative model of Gaussian LDA.

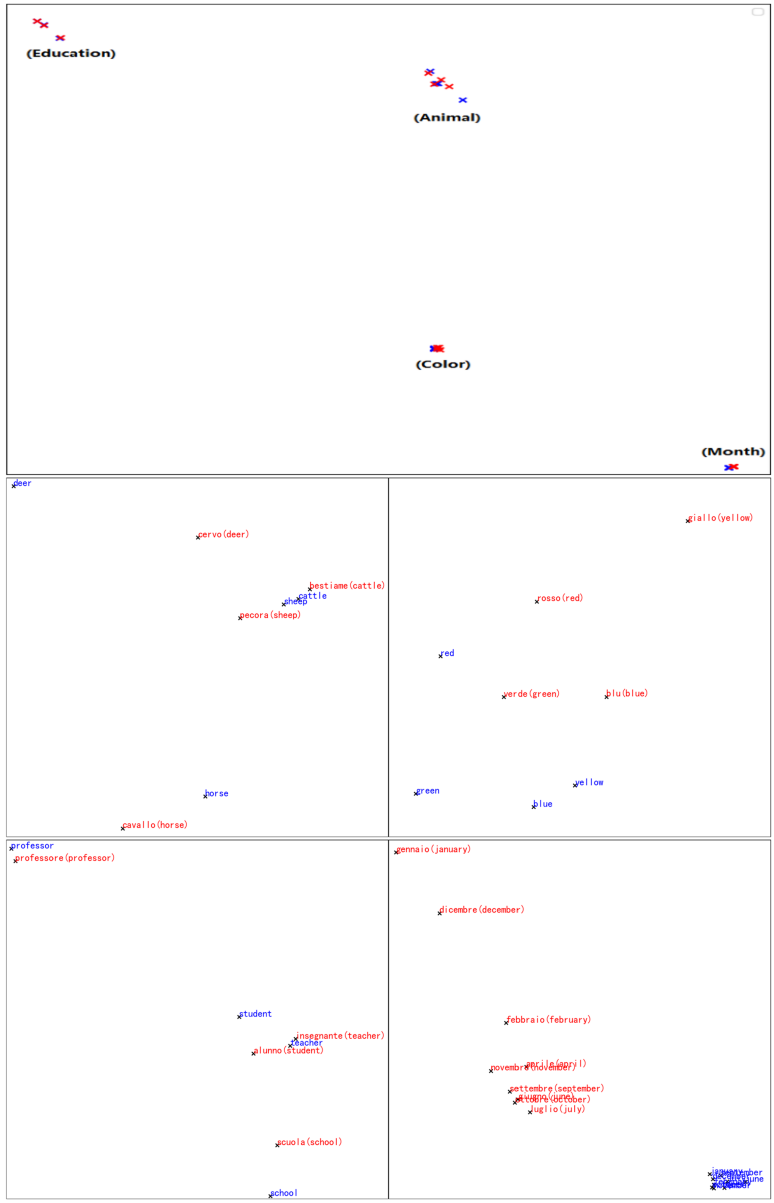

Fig. 2. Word embeddings from topics *animals*, *colors*, *education*, and *months* in English (blue) and Italian (red) after alignment by MUSE. In the top figure, all the embeddings are displayed together by markers, showing that the linear correspondence holds at a coarse level. In the following figures, the embeddings are zoomed in and displayed by topic (row-major order), only to show that the linear correspondence does not hold at local level. All the embeddings have been projected to two dimensions using t-SNE [26].

4.1.2 GMM. The GMM similarly assumes the words to be numerical embeddings, normally distributed within each topic. However, it dispenses with all the priors, making it possible to estimate all the parameters with a simple **expectation-maximization** (**EM**) approach rather than by variational inference or Monte Carlo sampling. Algorithm 2 shows the main steps of the EM parameter estimation for a GMM.

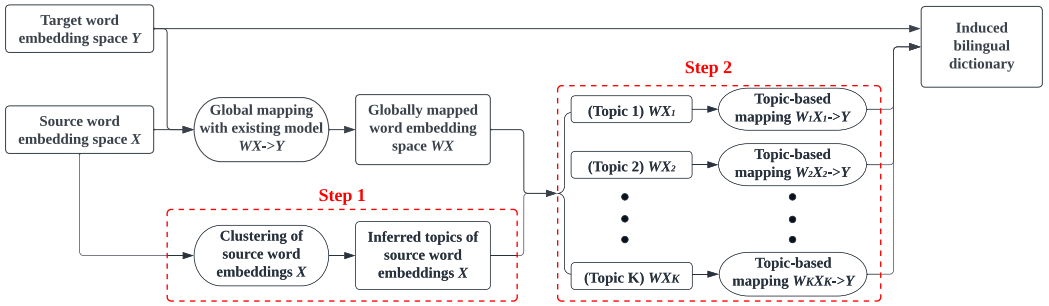

Fig. 3. Architecture diagram of the proposed approach.

ALGORITHM 1: Gaussian LDA: generative model.

for *topic* $k = 1 \ldots K$ **do**
  Draw covariance $\Sigma_k \sim W^{-1}(\Psi, \nu)$;
  Draw mean $\mu_k \sim N(\mu, \Sigma_k)$;
for *document* d *in corpus* D **do**
  Draw topic distribution $\theta_d \sim \text{Dir}(\alpha)$;
  for *word index* $n = 1 \ldots N_d$ **do**
    Draw a topic $z_n \sim \text{Categorical}(\theta_d)$;
    Draw a word embedding $w_{d,n} \sim N(\mu_{z_n}, \Sigma_{z_n})$;

ALGORITHM 2: Gaussian mixture model: expectation-maximization.

Initialize each Gaussian component, $N(\mu_k, \Sigma_k)$, with random parameters;
while *not converged* **do**
  for *each word embedding* w **do**
    for *component* $k = 1 \ldots K$ **do**
      Compute probability $N(w|\mu_k, \Sigma_k)$;
  for *component* $k = 1 \ldots K$ **do**
    Update its μ_k, Σ_k parameters using the computed probabilities as sample weights;

After the topics are inferred with either Gaussian LDA or the GMM, each word embedding of the source language is hard-assigned to the topic with highest probability.

4.2 Topic-Based Mapping

Once the word embeddings have been clustered, we first infer a global mapping of the source embedding space using MUSE (other global alignment approaches could also be employed), noting the corresponding mapping matrix as W_{global}. We then build a seed dictionary for each cluster to compute a set of K mapping matrices at cluster level. Specifically, we propose an adaptive dictionary construction to produce high-quality seed dictionaries that can effectively "refine" the mapping matrices at cluster level.

Similarly to MUSE, we limit the seed dictionaries to word pairs that are mutual nearest neighbors, and by using only the most-frequent words of the target language. However, since clustering breaks the source space into smaller sets, the main challenge is that some of the clusters may not

have enough mutual neighbors, or may simply be too small in the first place. In both these cases, the mapping matrix for the cluster, W_k, is just kept as $W_k = W_{global}$. Operationally, we build the seed dictionary incrementally: first, we set an initial threshold on the minimum acceptable frequency and search for mutual neighbors within this range. Then, if the size of this dictionary is still too small and there are more target words available, we relax the threshold on the minimum frequency and we repeat this step. As a final stage, we infer the mapping matrix for the cluster, W_k, using the current mapped source embeddings and the matched target embeddings as the arguments for the Procrustes optimization. To ensure the best possible mapping, we iterate this procedure N times by applying the matrix inferred at the previous iteration to the current mapped source embeddings, and picking the best mapping with a selection criterion. In the unsupervised case, similarly to MUSE, the selection criterion is the average CSLS distance of 10K matching pairs under the current W_k, while in the supervised case we simply use the accuracy over the provided supervised dictionary. Algorithm 3 shows the main steps of the the adaptive dictionary construction. If the size of the cluster, $size_k$, is less than a given threshold, γ, we immediately abandon the induction for that cluster. Otherwise, we begin an iterative process where we relax the threshold on the minimum frequency for the target word at every iteration. As in MUSE, the frequency of a word is expressed as the "rank" in its set; e.g., "1,000" means the word with the 1,000th frequency in the set in decreasing order. Therefore, parameter r sets the number of most-frequent words used for the induction. In turn, the size of the induced dictionary for the k-th cluster is noted as $size_{\theta_{k,r}}$, the size of the target set as $size_{target}$, and two tunable hyperparameters as α and β.

ALGORITHM 3: Adaptive construction of the clusters' seed dictionaries.

for *cluster* $k = 1 \ldots K$ **do**
  if $size_k < \gamma$ **then**
    Retain $W_k = W_{global}$;
  for *iteration* $i = 1 \ldots N$ **do**
    Initialize the rank parameter, r;
    Search for mutual neighbors and generate seed dictionary $\theta_{k,r}$;
    while $size_{\theta_{k,r}} < size_k * \alpha$ **and** $r < size_{target}$ **do**
      Relax the rank parameter, $r = r + \beta$;
      Search for mutual neighbors and update seed dictionary $\theta_{k,r}$;
    if $size_{\theta_{k,r}} < size_k * \alpha$ **then**
      Retain $W_k = W_{global}$;
    else
      Infer mapping matrix $W_k^{(i)} = \text{Procrustes}(\theta_{k,r})$;
  Select the best $W_k^{(i)}$;

In case a supervised dictionary is provided (the supervised case), we merge it with the constructed dictionary to form a larger seed dictionary, typically also of higher quality. However, many manually composed dictionaries, including those of the MUSE library,[1] are polysemic, i.e., may have multiple paired target words for a single source word. In this case, we only retain the closest target (based on the mapped word embeddings of the current iteration). In Section 5.3.3, we show the impact of this choice with respect to alternatives.

[1] https://github.com/facebookresearch/MUSE.

After we have learned the mapping matrices of all clusters, the final bilingual dictionary can be induced in two different ways: (1) by treating all the mapped source embeddings as a single set when performing the CSLS search (an approach that we refer to as "combined"); or (2) by searching for nearest neighbors using only a single cluster at a time, and then aggregating all the matched pairs ("separate"). On the whole, the first approach enjoys more freedom in matching, but may suffer from the structural differences between clusters. The second approach restricts the candidate set *a priori*, but may benefit from a better "linearized" mapped source space. In essence, the two approaches differ in how they compute the CSLS distance, and can lead to possibly significant differences in accuracy. In Section 5, we show that the differences in performance are in fact quite marked.

5 EXPERIMENTS

5.1 Experimental Settings

Languages. In the experiments, we focus on word translation between English (en) and other, diverse languages such as German (de), French (fr), Italian (it), Spanish (es), Finnish (fi), Turkish (tr), and Chinese (zh) (both directions), and between European languages, including German (de), French (fr), Italian (it), and Spanish (es) (all pairwise combinations), for a total of 26 language pairs.

Data. The pretrained word embeddings are from `fastText`[2] [5]. They are 300 dimensional and trained using the skip-gram model on Wikipedia corpora.[3] All words have been lower-cased and sorted according to their frequency. In order to fit within our GPU memory (6 GB), we have only used the 150K most frequent words for the experiments. In addition, all word embeddings have been zero-centered.

Parallel dictionaries. For the experiments, we have used the bilingual dictionaries from the `MUSE` library as ground truth. These dictionaries have been generated using an internal translation tool that can account for polysemic words (i.e., associate multiple target words to a single source word). For each language pair, they contain two dictionaries: a training set with 5,000 source words that can be used in the supervised case, and a test set with 1,500 source words. As in MUSE, to evaluate the accuracy (often also referred to as precision@1 in the literature), we count as correct matches those where the predicted word translation matches the ground-truth word translation (or any of them, in the case of multiple ground-truth translations), and then calculate the fraction of correct translations as a percentage. For example, if an algorithm predicts 1,100 correct translations out of the test set, its accuracy will be reported as $1{,}100/1{,}500 \approx 73.33\%$.

Unsupervised baseline. A thorough experimental comparison of word translation approaches was presented in [14]. The comparison has shown that MUSE, augmented with stochastic induction of the dictionary, has outperformed a number of contemporary approaches such as [15], [3], and [1], and can be regarded as the state-of-the-art approach for the field. Given the large number of language pairs and combinations, we have decided to organize our experimental results as a vis-à-vis comparison with MUSE, and add a more concise comparison with other approaches. We have also carried out preliminary experiments using MUSE augmented with the stochastic dictionary induction, but, with our settings, it did not show any improvements in accuracy over the base version. Therefore, we have not included the stochastic step in the baseline. However, inspired by it, we have also conducted experiments adding dropout during the inference of the mapping matrix. The full results are presented in Section 5.2.

[2] https://fasttext.cc/docs/en/pretrained-vectors.html.
[3] https://www.wikipedia.org/.

Supervised baseline. As the supervised baseline, we have used MUSE with the default parameter settings and a ground-truth seed dictionary of 5,000 source words. Results with different management of the polysemes are presented in Section 5.3.3.

Gaussian LDA and Gaussian mixture model. To implement the Gaussian LDA clustering, we have adapted the Multilingual Gaussian LDA code[4] of Kamyab [17]. The learning rate has been set to 0.05 as recommended, and the model has been trained for 20 iterations. For the Gaussian mixture model, we have used the Scikit Learn library.[5] For reproducibility, the seed for the random initialization (parameter `random_state`) has been set to a fixed value (42), and `max_iter` has been set to 200. The word embeddings have been clustered into 30 topics. To select this number, we have conducted a set of initial experiments, the results of which are reported in Section 5.3.2.

Hyperparameters. As hyperparameters for Algorithm 3, we have used minimum cluster size $\gamma = 15{,}000$, initial rank parameter $r = 15{,}000$, size ratio $\alpha = 1/3$, and rank increment $\beta = 5{,}000$. The two steps of mapping matrix inference and dictionary induction have been iterated $N = 5$ times. Again, these values have been chosen based on preliminary experiments.

Statistical significance. In order to test the statistical significance of the reported performance improvements, we have conducted a bootstrap test[6] as suggested in Dror et al. [11] on all our main results. The results that have passed the significance test (p-value < 0.05) have been marked with an asterisk (*) in Tables 1, 3, 4, and 5.

5.2 Main Results

In this section, we report our main experimental results, first for the unsupervised case and then for the supervised.

5.2.1 Unsupervised Case. To prepare a competitive baseline for each dataset, we have first trained MUSE three times and retained the model with the highest test-set accuracy. Our approach has been applied on top of this baseline.

Translations from other languages to English. Table 1 shows the results achieved by the proposed approach against the baseline in the translation toward English in the unsupervised case, with different clustering techniques (Gaussian LDA vs. GMM) and dictionary induction strategies ("combined" vs. "separate"). The relative improvements have been very marked in all cases, from a minimum of +0.07 pp to a maximum of +6.16 pp. The best average improvement across all languages (+3.30 pp) has been achieved with the Gaussian LDA clustering and the "separate" strategy. The languages that seem to have benefited the most are those that are the most "distant" from English, such as Finnish and Chinese. Conversely, the improvements from French to English have been the most limited. This is in good accordance with the intuition that farther languages would differ more in the structure of their word embeddings spaces; more evidence is provided in Subsection 5.3.1.

Comparison with other models. Table 2 shows a comparison of our results with those recently reported by Li et al. [18] for the languages in common. The results show that the proposed approach is, in general, very competitive.

Translations between European languages. In turn, Table 3 shows the results achieved in the translation between European languages. The improvements have been marked also in this case, with a best average across all pairs of +1.49 pp.

[4] https://github.com/EliasKB/Multilingual-Gaussian-Latent-Dirichlet-Allocation-MGLDA.
[5] https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html.
[6] https://github.com/rtmdrr/testSignificanceNLP.

Table 1. Test-Set Accuracy of the Induced Dictionaries in Translation from Other Languages to English in Unsupervised Settings

Languages	Baseline	Proposed Approach (Combined)	Improvement (pp)	Proposed Approach (Separate)	Improvement (pp)
		Gaussian LDA			
es-en	83.38	85.58	+2.20*	85.58	+2.20*
de-en	73.35	75.48	+2.14*	75.68	+2.34*
fr-en	82.36	82.77	+0.40	82.43	+0.07
tr-en	58.02	60.27	+2.25*	62.12	+4.10*
fi-en	59.12	61.66	+2.55*	64.81	+5.70*
it-en	78.44	80.37	+1.94*	80.97	+2.54*
zh-en	32.15	37.84	+5.69*	38.31	+6.16*
Average	66.69	69.14	+2.45	69.99	+3.30
		GMM			
es-en	83.38	86.05	+2.67*	85.85	+2.47*
de-en	73.35	75.75	+2.40*	75.89	+2.54*
fr-en	82.36	84.03	+1.67*	84.03	+1.67*
tr-en	58.02	63.28	+5.26*	62.73	+4.71*
fi-en	59.12	64.28	+5.16*	63.94	+4.83*
it-en	78.44	81.31	+2.87*	81.38	+2.94*
zh-en	32.15	33.82	+1.67*	33.15	+1.00
Average	66.69	69.79	+3.10	69.57	+2.88

"Combined" refers to inferring the matching pairs over the entire target dictionary, while "separate" refers to inferring within each cluster. Results marked with * have passed the significance test.

Table 2. A Comparison with Results Reported in the Literature for the Main Models

Approach	es-en	de-en	fr-en	it-en	zh-en
Conneau et al. [7]	83.66	72.93	82.39	77.90	32.33
Artetxe et al. [3]	84.60	74.27	83.60	79.80	0
Hartmann et al. [14]	84.73	74.20	83.73	79.67	34.53
Vulić et al. [28]	85.47	75.33	84.00	80.60	**44.07**
Li et al. [18]	84.60	74.33	83.67	79.67	35.93
Proposed approach (best)	**86.05**	**75.89**	**84.03**	**81.38**	38.31

5.2.2 Supervised Case. Table 4 shows the results for the translation toward English in the supervised case. The trend is slightly different from that of the unsupervised case, with the GMM clustering and the "separate" induction strategy always being better than the alternatives. The average improvement has been even higher than in the unsupervised case (+3.95 pp), with a maximum for the translation from Finnish to English (+6.97 pp).

In turn, the results for the translation between European languages are shown in Table 5. The largest average improvement has been +1.43 pp and has been obtained with (GMM + "combined"). However, the sensitivity to these hyperparameters has not been very pronounced, with the smallest average improvement (Gaussian LDA + "separate") still being +1.13 pp.

5.3 Analysis

In this section, we provide an analysis of various key aspects of the proposed approach, namely, (1) the clustered word embedding spaces; (2) the selection of the number of topics; (3) the management of the polysemes in the seed dictionary during supervised training; (4) a comparison

Table 3. Test-Set Accuracy of the Induced Dictionaries in Translation between European Languages in Unsupervised Settings

Languages	Baseline	Proposed Approach (Combined)	Improvement (pp)	Proposed Approach (Separate)	Improvement (pp)
		Gaussian LDA			
es-de	65.86	66.93	+1.07	67.07	+1.20*
es-fr	85.91	86.85	+0.93*	86.38	+0.47
es-it	83.23	84.03	+0.80*	83.43	+0.20
de-es	63.04	64.59	+1.54*	64.79	+1.74*
de-fr	66.98	71.54	+4.56*	70.81	+3.83*
de-it	69.17	70.52	+1.34*	70.18	+1.01*
fr-es	83.11	84.18	+1.07*	83.38	+0.27
fr-de	69.01	70.08	+1.07*	70.35	+1.34*
fr-it	81.31	82.44	+1.13*	81.44	+0.13
it-es	87.26	87.46	+0.20	86.86	−0.40
it-de	65.55	66.76	+1.21*	66.89	+1.34*
it-fr	88.26	88.13	−0.13	87.46	−0.80
Average	75.73	76.96	+1.23	76.59	+0.86
		GMM			
es-de	65.86	67.34	+1.47*	66.93	+1.07*
es-fr	85.91	87.25	+1.34*	86.78	+0.87*
es-it	83.23	83.97	+0.73	83.43	+0.20
de-es	63.04	65.33	+2.28*	64.86	+1.81*
de-fr	66.98	70.40	+3.42*	70.07	+3.09*
de-it	69.17	70.58	+1.41*	69.78	+0.60
fr-es	83.11	84.71	+1.60*	84.25	+1.13*
fr-de	69.01	70.82	+1.81*	70.21	+1.20*
fr-it	81.31	83.18	+1.87*	82.51	+1.20*
it-es	87.26	87.86	+0.60	87.59	+0.33
it-de	65.55	66.42	+0.87	66.29	+0.74*
it-fr	88.26	88.73	+0.47	88.39	+0.13
Average	75.73	77.21	+1.49	76.76	+1.03

Results marked with * have passed the significance test.

of the word matching step performed with the cosine similarity instead of CSLS; (5) experiments with dropout; (6) translation from English to other languages; (7) commented examples of word translations; and (8) the impact of topic modeling on the dictionary induction.

5.3.1 Analysis of the Clustered Embedding Spaces. Figure 4 shows the embeddings of Figure 1 after clustering with Gaussian LDA and the GMM, respectively (NB: each topic is rendered in a different color). At large, both Gaussian LDA and the GMM have been able to effectively cluster the word embeddings of both languages. However, for English the topics extracted with Gaussian LDA (top left) show a noticeable amount of overlap in the bottom-right part of the space. While this may depend on the parameters chosen for the experiments, or even just be an artifact of the t-SNE projection itself, we conclude that GMM may be better than Gaussian LDA in this case. Conversely, the clustering by Gaussian LDA of the Chinese word embeddings (top right) seems to be better than that of the GMM, as the colors are rarely mixed. This may offer a justification for their significant difference in accuracy (e.g., +6.16 pp vs. +1.00 pp in Table 1).

5.3.2 Selection of the Number of Topics. The number of topics is obviously an important parameter of the proposed approach. A larger number of topics makes the model more flexible; yet, at the same time, it makes the word clusters smaller and can, potentially, compromise the effectiveness of the generated seed dictionary. To illustrate this tradeoff, Table 6 shows the accuracies achieved by the proposed approach ("separate" + unsupervised) in translating from Finnish to English with

Table 4. Test-Set Accuracy of the Induced Dictionaries in Translation from Other Languages to English in Supervised Settings

Languages	Baseline	Proposed Approach (Combined)	Improvement (pp)	Proposed Approach (Separate)	Improvement (pp)
		Gaussian LDA			
es-en	84.11	86.52	+2.40*	86.78	+2.67*
de-en	73.68	76.69	+3.01*	77.42	+3.74*
fr-en	83.23	83.57	+0.33	84.24	+1.00*
tr-en	61.77	65.32	+3.55*	66.96	+5.19*
fi-en	60.99	63.94	+2.95*	66.89	+5.90*
it-en	77.84	81.38	+3.54*	82.04	+4.21*
zh-en	38.85	40.52	+1.67*	41.06	+2.21*
Average	68.64	71.13	+2.49	72.20	+3.56
		GMM			
es-en	84.11	86.38	+2.27*	86.85	+2.74*
de-en	73.68	76.95	+3.27*	78.22	+4.54*
fr-en	83.23	84.44	+1.20*	84.70	+1.47*
tr-en	61.77	65.87	+4.10*	65.87	+4.10*
fi-en	60.99	67.23	+6.23*	67.96	+6.97*
it-en	77.84	81.98	+4.14*	82.24	+4.41*
zh-en	38.85	41.06	+2.21*	42.26	+3.42*
Average	68.64	71.99	+3.35	72.59	+3.95

Results marked with * have passed the significance test.

different numbers of topics. The results show that setting the number of topics, K, to 30 has led to the best performance for both Gaussian LDA and the GMM. The variance has been much smaller for the GMM, which is therefore preferable on this account. In all cases, for this language pair the proposed approach has outperformed the baseline for any tested number of topics. For this reason, we have decided to not tune this parameter further and carried out all experiments with $K = 30$.

5.3.3 Management of Polysemes in the Seed Dictionary. The supervised training dictionaries provided with the MUSE library contain a substantial percentage of polysemes, which offers the opportunity to explore different ways to manage them in the supervised case. To this aim, we have conducted experiments with three different settings:

- **Setting 1:** the full supervised dictionary is only used in the first iteration. The following iterations only use the seed dictionary induced from the mutual nearest neighbors.
- **Setting 2:** The full supervised dictionary is used at every iteration. Starting with the second iteration, the seed dictionary is the combination of the supervised dictionary and that induced from the mutual nearest neighbors.
- **Setting 3:** This setting is similar to Setting 2, but before the combination, for each polyseme we only retain the translation with highest cosine similarity based on the word embeddings at the current iteration.

Table 7 shows the accuracies for the three different settings. Using the supervised dictionary at every iteration (Settings 2 and 3) has been consistently better than Setting 1, and retaining only one target word per polyseme (Setting 3) has led to a further improvement. This setting has been adopted in all our supervised experiments.

5.3.4 Comparison with the Cosine Similarity. The CSLS normalized distance has proved effective at mollifying the "hubness" of the word embedding spaces [7, 13]. However, we are concerned that the breaking up of the embedding space into smaller clusters performed by the proposed

Table 5. Test-Set Accuracy of the Induced Dictionaries in Translation from Other Between European Languages in Supervised Settings

Languages	Baseline	Proposed Approach (Combined)	Improvement (pp)	Proposed Approach (Separate)	Improvement (pp)
		Gaussian LDA			
es-de	70.55	71.69	+1.14*	73.16	+2.61*
es-fr	86.98	87.78	+0.80*	87.32	+0.33
es-it	83.90	85.77	+1.87*	84.90	+1.00*
de-es	69.08	70.42	+1.34*	70.62	+1.54*
de-fr	76.64	78.26	+1.61*	78.39	+1.74*
de-it	72.26	74.41	+2.15*	74.48	+2.22*
fr-es	83.18	84.45	+1.27*	84.25	+1.07*
fr-de	72.82	72.96	+0.13	73.29	+0.47
fr-it	82.98	84.18	+1.20*	83.51	+0.53*
it-es	87.39	88.26	+0.87*	88.06	+0.67*
it-de	70.79	72.33	+1.54*	72.46	+1.68*
it-fr	88.66	89.59	+0.93*	88.39	−0.27
Average	78.77	80.01	+1.24	79.90	+1.13
		GMM			
es-de	70.55	71.15	+0.60	71.62	+1.07*
es-fr	86.98	87.98	+1.00*	87.65	+0.67
es-it	83.90	85.50	+1.60*	84.64	+0.73*
de-es	69.08	70.42	+1.34*	70.36	+1.27*
de-fr	76.64	78.46	+1.81*	78.46	+1.81*
de-it	72.26	74.82	+2.55*	74.68	+2.42*
fr-es	83.18	84.85	+1.67*	84.91	+1.74*
fr-de	72.82	73.56	+0.74	73.56	+0.74
fr-it	82.98	84.45	+1.47*	84.05	+1.07*
it-es	87.39	89.06	+1.67*	89.13	+1.73*
it-de	70.79	72.60	+1.81*	72.73	+1.95*
it-fr	88.66	89.53	+0.87*	88.59	−0.07
Average	78.77	80.20	+1.43	80.03	+1.26

Results marked with * have passed the significance test.

Table 6. Accuracy of the Finnish-English Dictionary Induced by the Proposed Approach with Varying Number of Topics ("Separate" Generation Step, Unsupervised Settings)

Number of clusters	10	20	30	40	50
Gaussian LDA	59.38	63.67	64.81	64.28	63.54
GMM	63.20	63.81	63.94	62.71	62.73
Baseline	59.12				

approach may limit its effectiveness. For this reason, we have carried out an experiment replacing CSLS with the cosine similarity, and compared the accuracies and improvements. To this aim, Table 8 reports the results achieved in the translation toward English using the cosine similarity (unsupervised case, "separate" induction strategy). While the accuracies have been lower than those of CSLS, the improvements with the proposed approach have been even more substantial, with averages of +6.07 pp and +6.15 pp with Gaussian LDA and the GMM, respectively, and a largest improvement of 11.72 pp for Chinese to English. While the baseline has dropped its average accuracy by almost 4 pp with the cosine similarity in place of CSLS, the proposed approach

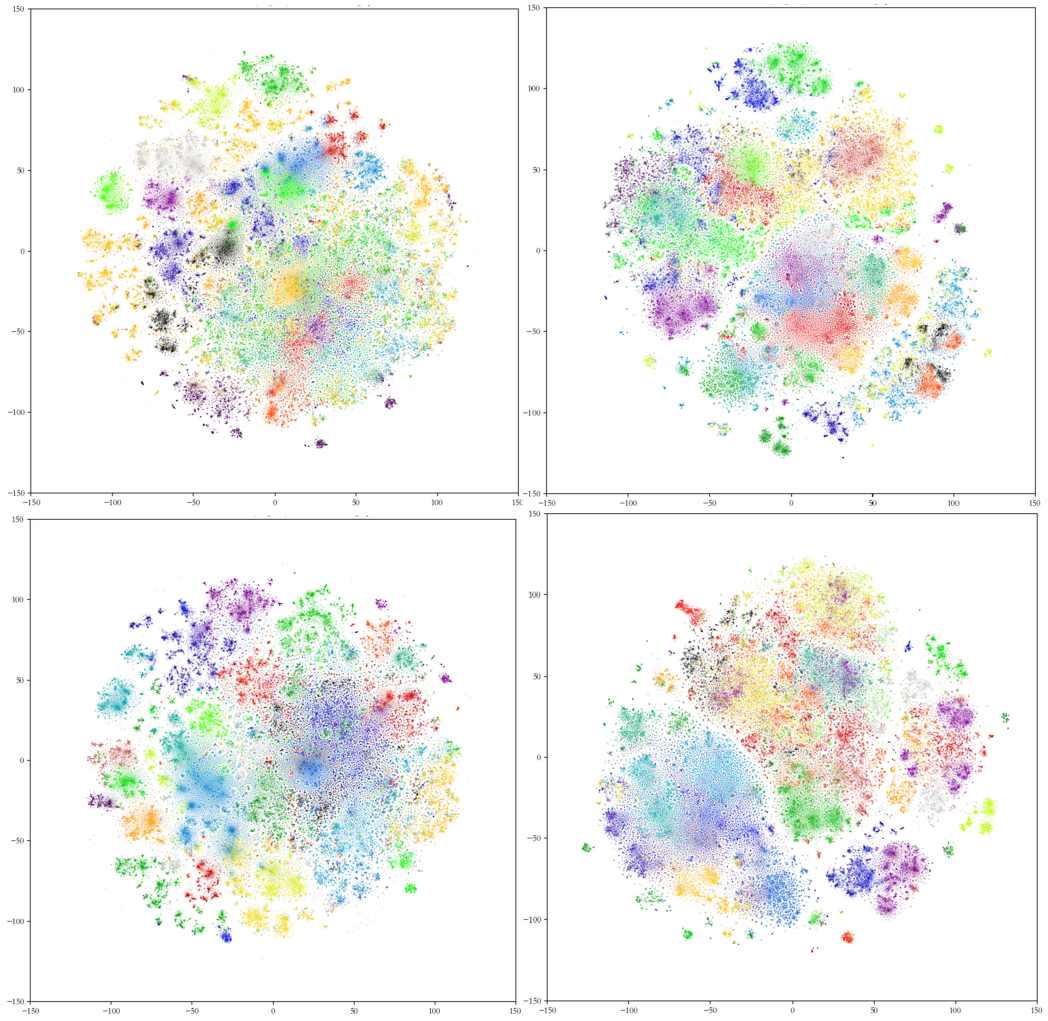

Fig. 4. The word embeddings for English (left) and Chinese (right) obtained with Gaussian LDA (top) and the GMM (bottom), clustered into 30 topics. Each topic is displayed with a different color. The embeddings have been projected to two dimensions using t-SNE [26].

has been able to perform almost on par. This is further evidence that the breaking up of the word embedding space into clusters leads to more accurate mapping matrices.

5.3.5 Experiments with Dropout. Inspired by the stochastic dictionary induction of [3], we have also conducted a few experiments adding a fixed dropout to the similarity matrix. As shown in Table 9, the performance of the baseline has mildly improved when setting the dropout to 0.1. However, the performance of our approach has been higher when applied to the baseline without dropout. A possible explanation is that the relatively small size of the clusters makes our approach more sensitive to the random removal of anchor points. Therefore, we have decided to retain the baseline without dropout for all our experiments.

Table 7. Accuracy of the Induced Dictionaries with Different Settings for the Management of Polysemes

Languages	Setting 1	Setting 2	Setting 3
es-en	83.18	83.71	84.11
de-en	73.81	73.21	73.68
fr-en	82.63	83.03	83.23
it-en	77.84	77.90	77.84
fi-en	59.65	59.99	60.99
tr-en	60.96	61.43	61.77
zh-en	38.85	39.12	38.85
Average	68.13	68.34	68.64

Table 8. Accuracy of the Induced Dictionaries using Nearest-Neighbor Matching (Translation to English, Unsupervised Case)

Languages	Baseline	Proposed Approach	Improvement (pp)	Proposed Approach	Improvement (pp)
		Gaussian LDA		GMM	
es-en	79.71	84.25	+4.54	84.71	+5.01
de-en	71.41	74.42	+3.01	75.35	+3.94
fr-en	79.43	82.16	+2.74	82.90	+3.47
tr-en	55.09	61.91	+6.83	62.32	+7.24
fi-en	56.37	64.28	+7.91	63.94	+7.57
it-en	74.77	80.51	+5.74	81.44	+6.68
zh-en	22.97	34.70	+11.72	32.15	+9.18
Average	62.82	68.89	+6.07	68.97	+6.15
Average (CSLS)	*66.69*	*69.99*	*+3.30*	*69.57*	*+2.88*

Table 9. Accuracy of the Induced Dictionaries for the Unsupervised Baseline with and without Dropout, and the Corresponding Accuracy with the Proposed Approach (tr-en and fi-en)

Approaches	tr-en	fi-en
Baseline	58.02	59.12
Proposed Approach	63.28	64.28
Baseline (dropout = 0.1)	58.84	59.72
Proposed Approach	61.98	63.67

5.3.6 Translation from English to other Languages. Tables 10 and 11 show the results achieved by the proposed approach in translation from English to other languages. In general, the improvements have been less pronounced than in the opposite direction, and we speculate that this may be due to the likely more irregular structure of the target word embeddings compared to the English word embeddings. In the unsupervised case, the GMM has performed remarkably better than Gaussian LDA, and the "combined" induction strategy may have been able to partially compensate for the irregular structure of the target space. In the supervised case, thanks to the more accurate guidance on the clusters, the "separate" strategy has performed the best, and Gaussian LDA has managed to achieve an average improvement of +1.82 pp.

Table 10. Test-Set Accuracy of the Induced Dictionaries in Translation from English to Other Languages in Unsupervised Settings

Languages	Baseline	Proposed Approach (Combined)	Improvement (pp)	Proposed Approach (Separate)	Improvement (pp)
		Gaussian LDA			
en-es	82.64	83.38	+0.73	82.98	+0.33
en-de	75.65	75.85	+0.20	76.32	+0.67
en-fr	82.86	81.99	−0.87	81.92	−0.93
en-tr	48.59	49.06	+0.47	48.99	+0.40
en-fi	45.66	47.07	+1.41	47.47	+1.82
en-it	78.10	78.64	+0.53	78.77	+0.67
en-zh	32.22	30.94	−1.27	30.68	−1.54
Average	63.67	63.85	+0.17	63.88	+0.20
		GMM			
en-es	82.64	83.78	+1.13	83.18	+0.53
en-de	75.65	76.98	+1.33	77.18	+1.53
en-fr	82.86	83.19	+0.33	83.26	+0.40
en-tr	48.59	50.74	+2.15	50.34	+1.74
en-fi	45.66	47.47	+1.82	47.41	+1.75
en-it	78.10	80.04	+1.94	79.57	+1.47
en-zh	32.22	31.08	−1.14	31.01	−1.21
Average	63.67	64.75	+1.08	64.56	+0.89

Table 11. Test-Set Accuracy of the Induced Dictionaries in Translation from English to Other Languages in Supervised Settings

Languages	Baseline	Proposed Approach (Combined)	Improvement (pp)	Proposed Approach (Separate)	Improvement (pp)
		Gaussian LDA			
en-es	82.71	83.91	+1.20	84.45	+1.74
en-de	76.12	77.12	+1.00	77.79	+1.67
en-fr	82.79	82.92	+0.13	83.06	+0.27
en-tr	49.87	51.74	+1.88	52.35	+2.48
en-fi	53.27	56.09	+2.83	57.17	+3.91
en-it	78.50	79.64	+1.13	79.71	+1.20
en-zh	43.40	42.80	−0.60	43.87	+0.47
Average	66.67	67.75	+1.08	68.34	+1.68
		GMM			
en-es	82.71	83.85	+1.13	84.05	+1.34
en-de	76.12	76.92	+0.80	76.72	+0.60
en-fr	82.79	83.46	+0.67	83.66	+0.87
en-tr	49.87	53.76	+3.89	54.30	+4.43
en-fi	53.27	55.56	+2.29	55.96	+2.69
en-it	78.50	80.91	+2.40	80.57	+2.07
en-zh	43.40	43.74	+0.33	44.14	+0.74
Average	66.67	68.31	+1.65	68.48	+1.82

5.3.7 Examples of Word Translations. As a qualitative analysis, Table 12 shows a few examples of word translations from Chinese to English for the baseline and the proposed approach in the unsupervised case (Gaussian LDA + "separate"). In some cases, the proposed approach has been able to correct the baseline, and vice versa. Some of the corrections performed by the proposed approach have been mild (e.g., "nightmare" → "mirage"; "capsicum" → "pepper") and they are probably thanks to its finer mapping. However, in other cases they have recovered from "catastrophic" translations (e.g., "reintroduced" → "bison"; "stating" → "ark"). On the other hand, at times the

Table 12. Examples of Chinese-English Word Translations using MUSE and the Proposed Approach

Source	MUSE	Proposed Approach	Ground Truth
Translated correctly by the **proposed approach (Gaussian LDA)**, but incorrectly by MUSE			
幻影	nightmare	**mirage**	phantom,mirage
焊接	plating	**welded**	weld,welded,welding
慢性	exacerbation	**chronic**	chronic
胡椒	capsicum	**pepper**	pepper
方舟	stating	**ark**	ark
野牛	reintroduced	**bison**	buffalo,bison
骨架	structurally	**skeleton**	skeleton,framing
色素	melanin	**pigment**	pigment
擦音	allophonic	**fricative**	fricative
河床	deltas	**riverbed**	riverbed
Translated correctly by **MUSE**, but incorrectly by the proposed approach (Gaussian LDA)			
牛肉	**beef**	pork	beef
山麓	**foothills**	slopes	piedmont,foothills
乙醇	**ethanol**	methanol	ethanol
Translated correctly by both the **proposed approach (Gaussian LDA)** and **MUSE**, but differently			
潮汐	**tides**	**tidal**	tide,tidal,tides
室外	**outdoor**	**outdoors**	outside,outdoor,outdoors
酸性	**acidity**	**acidic**	acidic,acid,acidity

proposed approach has introduced its own errors: mild in some cases ("slopes" for "foothills"), and major in others ("pork" for "beef"). These latter errors are possibly due to the side effects of clustering and the "separate" induction strategy. In addition, we report a few cases where the translations have been different, yet correct for both approaches, given the multiple, allowable targets (polysemes).

5.3.8 Impact of the Topic Modeling on the Dictionary Induction. In this subsection, we analyze the impact of the topic modeling on the dictionary induction, with emphasis on the mapping of the source word embeddings to the target space. The main indicators for the Chinese to English case (Gaussian LDA, unsupervised) are reported in Table 13.

Topic overview. To give an idea of each topic's content, the column *Words* in Table 13 lists the topic's five most-frequent words, translated into English for easier comprehension. Hereafter is also a brief description of the topics that we were able to connote:

Topic 2: mainly symbols and single Chinese characters
Topic 5: mainly countries and cities (Simplified Chinese)
Topic 7: mainly words related to Hong Kong and Taiwan (Traditional Chinese)
Topic 8, 21: mainly words related to Japanese
Topic 12: mainly countries and cities (Traditional Chinese)
Topic 16: mainly words related to eras, ages, periods
Topic 18: mainly words related to fishing
Topic 22: mainly adverbs and idioms (Simplified Chinese)
Topic 27: mainly adverbs and idioms (Traditional Chinese)

Impact of the proposed topic-based approach. In Table 13, the column *Count* shows the word count (i.e., size) of each topic in the source language, and the column *Test* shows the word

Table 13. Detailed Comparison by Topic of the Baseline Model (Base) and the Proposed Approach (Prop) for Dictionary Induction from Chinese to English (Gaussian LDA, Unsupervised)

Topic	Words	Count	Test	Acc (Base)	Acc (Prop)	Cos (Base)	Cos (Prop)
1	series, you, sun, center, chestnut	2273	0	0.0	0.0	0.4568	0.6472
2	*single Chinese Character*	5974	130	0.0	-	0.8912	-
3	name, wikidata, original, team, border	2323	2	0.0	-	0.6488	-
4	one, orientation, research, community, similar	16086	270	37.41	44.07	0.5039	0.5522
5	Chinese, America, Beijing, French, Italy	4115	35	51.43	62.86	0.5011	0.5825
6	this, and, is, all, not	4064	0	0.0	-	0.4584	-
7	Hong Kong, Taiwan, Chinese, commence, university	5706	184	33.15	34.78	0.5448	0.6088
8	Japan, Tokyo, Showa, write lyrics, starring	8311	29	6.9	-	0.4903	-
9	use, in, and, mainly, include	3491	95	38.95	36.84	0.5829	0.6365
10	one, become, because, not, due to	4763	93	29.03	34.41	0.5676	0.6434
11	center, image, title, group, file	3714	0	0.0	-	0.4633	-
12	America, Republic of China, TV station, French, Europe	5735	117	46.15	61.54	0.4869	0.5759
13	pt, sg, sw, fn, sh	3078	0	0.0	0.0	0.4127	0.5632
14	lie in, time, programme, play, sports	12353	327	38.23	43.43	0.5040	0.5632
15	space, gravity field, observatory, laser, measuring instrument	6740	49	46.94	53.06	0.4736	0.5667
16	born, twenty, era, Qing Dynasty, empire	6736	46	19.57	-	0.5138	-
17	handwriting, default, lifeline, brag, straight	6180	1	0.0	0.0	0.5569	0.6311
18	fish, net, fin, carp, catfish	4435	3	0.0	-	0.8560	-
19	role, song, story, show, host	5442	94	40.43	46.81	0.5364	0.5993
20	great, Mario, Spain, building, high	3345	0	0.0	0.0	0.4858	0.6456
21	Hokkaido, railway, Japanese Emperor, chome, East Japan	2162	1	0.0	0.0	0.4234	0.5903
22	first time, mainly, in the world, excusable, especially	2472	11	36.36	36.36	0.4331	0.6371
23	new, love, one, time, black	6092	0	0.0	0.0	0.3764	0.5866
24	music, Windows, world, Google, Facebook	4967	0	0.0	0.0	0.3503	0.6331
25	John, David, James, Robert, Michael	4053	0	0.0	-	0.3420	-
26	Shanghai, India, Dan, Taiwan, Malaysia	3509	0	0.0	0.0	0.4246	0.6151
27	not at all, actually, sickly, in fact, influence	2698	6	66.67	50.0	0.4198	0.6374
28	program, system, analysis, control, science	3075	0	0.0	-	0.4021	-
29	holy, last, from, we, of	3543	0	0.0	-	0.4891	-
30	genus, species, class, animal, family	2565	0	0.0	-	0.4451	-

count of each topic in the test set (1,500 words in total). The rest of the table illustrates the impact of the proposed topic-based approach on the dictionary induction and the mapping. To this aim, the column *Acc (Base)* reports the accuracy of the MUSE baseline by topic, while the column *Acc (Prop)* reports that of the proposed approach. In turn, the column *Cos (Base)* reports the average cosine similarity between the mapped source words and their closest target word, and the column *Cos (Prop)* shows the same figure for the proposed approach. Cells marked with a "-" correspond to topics for which no mapping matrix was computed (see Algorithm 3).

The most immediate observation from Table 13 is that the cosine similarity between the mapped source words and the target words has been consistently increased by the proposed approach. This shows the good generalization of the mappings learned from the constructed seed dictionaries. In addition, the size filters in Algorithm 3 have prevented learning mappings for topics that, in most cases, have a very small word count in the test set (except topic 2). Finally, the table shows a good correspondence between most of the increases in cosine similarity and the eventual increases in accuracy (with the notable exception of topic 27, where the accuracy has dropped by 16.67 pp). Overall, these results seem to confirm the effectiveness of the proposed approach.

6 CONCLUSION

This article has proposed a novel approach for bilingual dictionary induction that first clusters the source word embeddings according to a topic model, and then infers a separate mapping matrix for each cluster. The seed dictionaries for inferring the mapping matrices are generated with an algorithm that carefully accounts for the size of the individual clusters and the number of matched pairs. A large range of experimental results in both unsupervised and supervised settings over 26 language pairs have shown that the proposed approach has been able to substantially

outperform a state-of-the-art approach in translation between English and languages such as German, French, Italian, Spanish, Finnish, Turkish, and Chinese, and between a set of European languages. The average improvements have reached +3.30 pp in the unsupervised case and +3.95 pp in the supervised case, and the results have also proved competitive against those of other recent models from the literature (Table 2). Overall, we believe that the main finding of our article is that the mapping at cluster level is a viable approach for managing the differences in structure between the source and target word embedding spaces, and could potentially be applied to other word translation approaches. In the future, we plan to explore other embedding distances that may better suit the proposed topic-based approach.

REFERENCES

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

[2] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 451–462. https://doi.org/10.18653/v1/P17-1042

[3] 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 (Volume 1: Long Papers)*. 789–798.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3 (2003), 993–1022.

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

[6] Hailong Cao and Tiejun Zhao. 2021. Word embedding transformation for robust unsupervised bilingual lexicon induction. arXiv:2105.12297. https://arxiv.org/abs/2105.12297.

[7] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. arXiv:1710.04087. https://arxiv.org/abs/1710.04087.

[8] Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 795–804. https://doi.org/10.3115/v1/P15-1077

[9] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics* 8 (2020), 439–453.

[10] Georgiana Dinu and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. *CoRR* abs/1412.6568 (2015). arXiv:1412.6568. https://arxiv.org/abs/1412.6568.

[11] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1383–1392. https://doi.org/10.18653/v1/P18-1128

[12] I. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial networks. arXiv:1406.2661. https://arxiv.org/abs/1406.2661.

[13] Edouard Grave, Armand Joulin, and Quentin Berthet. 2018. Unsupervised alignment of embeddings with Wasserstein procrustes. *CoRR* abs/1805.11222 (2018). arXiv:1805.11222. http://arxiv.org/abs/1805.11222.

[14] Mareike Hartmann, Yova Kementchedjhieva, and Anders Søgaard. 2019. Comparing Unsupervised Word Translation Methods Step by Step. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS'19)*. 6031–6041.

[15] Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 469–478. https://doi.org/10.18653/v1/d18-1043

[16] Sanjanasri J. P., Vijay Krishna Menon, Soman K. P., Rajendran S., and Agnieszka Wolk. 2021. Generation of cross-lingual word vectors for low-resourced languages using deep learning and topological metrics in a data-efficient way. *Electronics* 10, 12 (2021), 1372:1–23.

[17] Elias Kamyab. 2019. *Multilingual Gaussian Latent Dirichlet Allocation.* Master's thesis. Chalmers University of Technology and University of Gothenburg, Sweden.

[18] Yanyang Li, Yingfeng Luo, Ye Lin, Quan Du, Huizhen Wang, Shujian Huang, Tong Xiao, and Jingbo Zhu. 2020. A simple and effective approach to robust unsupervised bilingual dictionary induction. In *Proceedings of the 28th International Conference on Computational Linguistics*.

[19] G. J. McLachlan and D. Peel. 2000. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley, New York.

[20] 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*. Association for Computational Linguistics, 121–126. https://doi.org/10.18653/v1/W16-1614

[21] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. arXiv:1309.4168. https://arxiv.org/abs/1309.4168.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13)*. Curran Associates Inc., Red Hook, NY, 3111–3119.

[23] 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'14)*. Association for Computational Linguistics, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[24] Peter Schönemann. 1966. A generalized solution of the orthogonal Procrustes problem. *Psychometrika* 31, 1 (1966), 1–10. https://EconPapers.repec.org/RePEc:spr:psycho:v:31:y:1966:i:1:p:1-10.

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

[26] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html.

[27] Ivan Vulić and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 247–257. https://doi.org/10.18653/v1/P16-1024

[28] Ivan Vulić, Anna Korhonen, and Goran Glavaš. 2020. Improving bilingual lexicon induction with unsupervised post-processing of monolingual word vector spaces. In *Proceedings of the 5th Workshop on Representation Learning for NLP*. 45–54.

[29] Ivan Vulić and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research* 55, 1 (Jan. 2016), 953–994.

[30] 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*. Association for Computational Linguistics, 1006–1011. https://doi.org/10.3115/v1/N15-1104

[31] Deyi Xiong, Fandong Meng, and Qun Liu. 2016. Topic-based term translation models for statistical machine translation. *Artificial Intelligence* 232 (2016), 54–75. https://doi.org/10.1016/j.artint.2015.12.002

[32] Deyi Xiong and Min Zhang. 2013. A topic-based coherence model for statistical machine translation. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI'13)*. AAAI Press, 977–983.

[33] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1959–1970. https://doi.org/10.18653/v1/P17-1179

Received 12 October 2021; revised 25 May 2022; accepted 14 September 2022