Aligning Vector-spaces with Noisy Supervised Lexicons

Noa Yehezkel Lubin[†1] **Jacob Goldberger**[‡2] **Yoav Goldberg**[†*3]

[†]Computer Science Department, Bar Ilan University, Ramat Gan, Israel
[‡]Electrical Engineering Department, Bar Ilan University, Ramat Gan, Israel
[*]Allen Institute for Artificial Intelligence

[1]noa.kel@gmail.com
[2]jacob.goldberger@biu.ac.il
[3]yoav.goldberg@gmail.com

Abstract

The problem of learning to translate between two vector spaces given a set of aligned points arises in several application areas of NLP. Current solutions assume that the lexicon which defines the alignment pairs is noise-free. We consider the case where the set of aligned points is allowed to contain an amount of noise, in the form of incorrect lexicon pairs and show that this arises in practice by analyzing the edited dictionaries after the cleaning process. We demonstrate that such noise substantially degrades the accuracy of the learned translation when using current methods. We propose a model that accounts for noisy pairs. This is achieved by introducing a generative model with a compatible iterative EM algorithm. The algorithm jointly learns the noise level in the lexicon, finds the set of noisy pairs, and learns the mapping between the spaces. We demonstrate the effectiveness of our proposed algorithm on two alignment problems: bilingual word embedding translation, and mapping between diachronic embedding spaces for recovering the semantic shifts of words across time periods.

1 Introduction

We consider the problem of mapping between points in different vector spaces. This problem has prominent applications in natural language processing (NLP). Some examples are creating bilingual word lexicons (Mikolov et al., 2013), machine translation (Artetxe et al., 2016, 2017a,b, 2018a,b; Conneau et al., 2017), hypernym generation (Yamane et al., 2016), diachronic embeddings alignment (Hamilton et al., 2016) and domain adaptation (Barnes et al., 2018). In all these examples one is given word embeddings in two different vector spaces, and needs to learn a mapping from one to the other.

The problem is traditionally posed as a supervised learning problem, in which we are given two sets of vectors (e.g.: word-vectors in Italian and in English) and a *lexicon* mapping the points between the two sets (known word-translation pairs). Our goal is to learn a mapping that will correctly map the vectors in one space (e.g.: English word embeddings) to their known corresponding vectors in the other (e.g.: Italian word embeddings). The mapping will then be used to translate vectors for which the correspondence is unknown. This setup was popularized by Mikolov et al. (2013).

The supervised setup assumes a perfect lexicon. Here, we consider what happens in the presence of *training noise*, where some of the lexicon's entries are incorrect in the sense that they don't reflect an optimal correspondence between the word vectors.

2 Background

2.1 The Supervised Translation Problem

We are given two datasets, $X = x_1, ..., x_m$ and $Y = y_1, ..., y_n$, coming from d-dimensional spaces $\mathcal{X}$ and $\mathcal{Y}$. We assume that the spaces are related, in the sense that there is a function $f(x)$ mapping points in space $\mathcal{X}$ to points in space $\mathcal{Y}$. In this work, we focus on linear mappings, i.e. a $d \times d$ matrix Q mapping points via $y_i = Qx_i$. The goal of the learning is to find the translation matrix Q. In the supervised setting, $m = n$ and we assume that $\forall i \ f(x_i) \approx y_i$. We refer to the sets X and Y as the *supervision*. The goal is to learn a matrix $\hat{Q}$ such the Frobenius norm is minimized:

$$\hat{Q} = \arg\min_{Q} \|QX - Y\|_F^2. \quad (1)$$

2.2 Existing Solution Methods

Gradient-based The objective in (1) is convex, and can be solved via least-squares method or via stochastic gradient optimization iterating over the

Proceedings of NAACL-HLT 2019, pages 460–465
Minneapolis, Minnesota, June 2 - June 7, 2019.

pairs (x_i, y_i), as done by Mikolov et al. (2013) and Dinu and Baroni (2014).

Orthogonal Procrustes (OP) Artetxe et al. (2016) and Smith et al. (2017) argued and proved that a linear mapping between sub-spaces must be orthogonal. This leads to the modified objective:

$$\hat{Q} = \underset{Q,s.t:Q^TQ=I}{\arg\min} \|QX - Y\|_F^2 \qquad (2)$$

Objective (2) is known as the *Orthogonal Procrustes Problem*. It can be solved algebraically by using a singular value decomposition (SVD). Schnemann (1966) proved that the solution to 2 is: $\hat{Q} = UV^T$ s.t. $U\Sigma V^T$ is the SVD of YX^T.

The OP method is used in Xing et al. (2015); Artetxe et al. (2016, 2017a,b, 2018a,b); Hamilton et al. (2016); Conneau et al. (2017); Ruder et al. (2018).

2.3 The Unsupervised Translation Problem

The supervised alignment problem can be expended to the semi-supervised (Artetxe et al., 2017b; Lample et al., 2017; Ruder et al., 2018) or unsupervised (Zhang et al., 2017; Conneau et al., 2017; Artetxe et al., 2018b; Xu et al., 2018; Alvarez-Melis and Jaakkola, 2018) case, where a very small lexicon or none at all is given. In iterative methods, the lexicon is expended and used to learn the alignment, later the alignment is used to predict the lexicon for the next iteration and so on. In adversarial methods, a final iterative step is used after the lexicon is built to refine the result. We will focus on the supervised stage in the unsupervised setting, meaning estimating the alignment once a lexicon is induced.

3 The Effect of Noise

The previous methods assume the supervision set X, Y is perfectly correct. However, this is often not the case in practice. We consider the case where a percentage p of the pairs in the supervision set are "noisy": applying the gold transformation to a noisy point x_j will not result in a vector close to y_j. The importance of the quality of word-pairs selection was previously analyzed by Vulić and Korhonen (2016). Here, we equate "bad pairs" to noise, and explore the performance in the presence of noise by conducting a series of synthetic experiments. We take a set of points X, a random transformation Q and a gold set $Y = QX$. We define *error* as $\|Y - \hat{Y}\|_F^2$ where $\hat{Y} = \hat{Q}X$ is the prediction according to the learned transform $\hat{Q}$. Following the claim that linear transformations between word vector spaces are orthogonal, we focus here on orthogonal transformations.

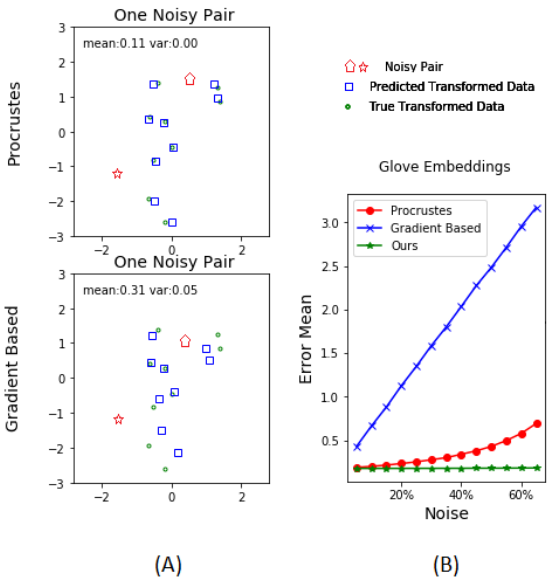

Figure 1: Noise influence. (A): the effect of a noisy pair on 2D alignment. (B) mean error over non-noisy pairs as a function of noise level.

Low Dimensional Synthetic Data We begin by inspecting a case of few 2-dimensional points, which can be easily visualized. We compare a noise-free training to the case of a single noisy point. We construct X by sampling $n = 10$ points of dimension $d = 2$ from a normal distribution. We take nine points and transformed them via an orthogonal random transform Q. We then add a single noisy pair which is generated by sampling two normally distributed random points and treating them as a pair. The error is measured only on the nine aligned pairs.

When no noise is applied, both Gradient-based and Procrustes methods are aligned with 0 error mean and variance. Once the noisy condition is applied this is no longer the case. Figure 1(A) shows the noisy condition. Here, the red point (true) and box (prediction) represent the noisy point. Green dots are the true locations after transformation, and the blue boxes are the predicted ones after transformation. Both methods are affected by the noisy sample: all ten points fall away from their true location. The effect is especially severe for the gradient-based methods.

High Dimensional Embeddings The experiment setup is as before, but instead of a normal distribution we use (6B, 300d) English Glove Embeddings (Pennington et al., 2014) with lexicon of size $n = 5000$. We report the mean error for various noise levels on an unseen aligned test set of size 1500.

In Figure 1(B) we can see that both methods are effected by noise. As expected, as the amount of noise increases the error on the test set increases. We can again see that the effect is worse with gradient-based methods.

4 Noise-aware Model

Having verified that noise in the supervision severely influences the solution of both methods, we turn to proposing a noise-aware model.

The proposed model jointly identifies noisy pairs in the supervision set and learns a translation which ignores the noisy points. Identifying the point helps to clean the underlying lexicon (dictionary) that created the supervision. In addition, by removing those points our model learns a better translation matrix.

Generative Model We are given $x \in \mathbb{R}^d$ and we sample a corresponding $y \in \mathbb{R}^d$ by first sampling a Bernoulli random variable with probability α:

$$z \sim \text{Bernoulli}(\alpha)$$

$$y \sim \begin{cases} N(\mu_y, \sigma_y^2 I) & z = 0\,(\text{`noise'}) \\ N(Qx, \sigma^2 I) & z = 1\,(\text{`aligned'}) \end{cases}$$

The density function y is a mixture of two Gaussians:

$$f(y|x) = (1-\alpha)N(\mu_y, {\sigma_y}^2 I) + \alpha N(Qx, \sigma^2 I).$$

The likelihood function is:

$$L(Q, \sigma, \mu_y, \sigma_y) = \sum_t \log f(y_t|x_t)$$

EM Algorithm We apply the EM algorithm (Dempster et al., 1977) to maximize the objective in the presence of latent variables. The algorithm has both soft and hard decision variants. We used the hard decision one which we find more natural, and note that the posterior probability of z_t was close to 0 or 1 also in the soft-decision case.

It is important to properly initialize the EM algorithm to avoid convergence to a local optima. We initialize Q by applying OP on the entire lexicon (not just the clean pairs). We initialize the variance, σ, by calculating $\sigma^2 = \frac{1}{n \cdot d} \sum_{t=1} \|Qx_t - y_t\|^2$. We initialize, μ_y, σ_y by taking the mean and variance of the entire dataset. Finally, we initialize α to 0.5.

The (hard version) EM algorithm is shown in Algorithm box 1. The runtime of each iteration is dominated by the OP algorithm (matrix multiplication and SVD on a $d \times d$ matrix). Each iteration contains an additional matrix multiplication and few simple vector operations. Figure 1(B) shows it obtains perfect results on the simulated noisy data.

Algorithm 1 Noise-aware Alignment

Data: List of paired vectors: $(x_1, y_1), ..., (x_n, y_n)$
Result: Q, σ, μ_y, σ_y
while $|\alpha_{curr} - \alpha_{prev}| > \epsilon$ **do**
  E step:
  $w_t = p(z_t{=}1|x_t, y_t) = \frac{\alpha N(Qx_t, \sigma^2 I)}{f(y_t|x_t)}$
  $h_t = 1(w_t > 0.5)$
  $n_1 = \sum_t h_t$
  M step:
  Apply OP on the subset $\{t|h_t = 1\}$ to find Q.
  $\sigma^2 = \frac{1}{d \cdot n_1} \sum_{t|h_t=1} \|Qx_t - y_t\|^2$
  $\mu_y = \frac{1}{(n-n_1)} \sum_{t|h_t=0} y_t$
  $\sigma_y^2 = \frac{1}{d(n-n_1)} \sum_{t|h_t=0} \|\mu_y - y_t\|^2$
  $\alpha_{prev} = \alpha_{curr}$
  $\alpha_{curr} = \frac{n_1}{n}$
end

5 Experiments

5.1 Bilingual Word Embedding

Experiment Setup This experiment tests the noise-aware solution on an unsupervised translation problem. The goal is to learn the "translation matrix", which is a transformation matrix between two languages by building a dictionary. We can treat the unsupervised setup after retrieving a lexicon as an iterative supervised setup where some of the lexicon pairs are noisy. We assumes the unsupervised setting will contain higher amount of noise than the supervised one, especially in the first iterations. We follow the experiment setup in Artetxe et al. (2018b). But instead of using OP for learning the translation matrix, we used our Noise-Aware Alignment (NAA), meaning we jointly learn to align and to ignore the noisy pairs. We used the En-It dataset provided by Dinu and Baroni (2014) and the extensions: En-De, En-Fi and En-Es of Artetxe et al. (2018a, 2017b).

Method	En→It			En→De			En→Fi			En→Es		
	best	avg	iters	best	avg	iters	best	avg	iters	best	avg	iters
Artetxe et al., 2018b	**48.53**	48.13	573	48.47	48.19	773	33.50	32.63	988	37.60	37.33	808
Noise-aware Alignment	**48.53**	**48.20**	471	**49.67**	**48.89**	568	**33.98**	**33.68**	502	**38.40**	**37.79**	551

Table 1: Bilingual Experiment P@1. Numbers are based on 10 runs of each method. The En→De, En→Fi and En→Es improvements are significant at $p < 0.05$ according to ANOVA on the different runs.

Experiment Results In Table 1 we report the best and average precision@1 scores and the average number of iterations among 10 experiments, for different language translations. Our model improves the results in the translation tasks. In most setups our average case is better than the former best case. In addition, the noise-aware model is more stable and therefore requires fewer iterations to converge. The accuracy improvements are small but consistent, and we note that we consider them as a lower-bound on the actual improvements as the current test set comes from the same distribution of the training set, and also contains similarly noisy pairs. Using the soft-EM version results in similar results, but takes roughly 15% more iterations to converge.

Table 2 lists examples of pairs that were kept and discarded in En-It dictionary. The algorithm learned the pair (dog → dog) is an error. Another example is the translation (good → santo) which is a less-popular word-sense than (good → buon / buona). When analyzing the En-It cleaned dictionary we see the percentage of potentially misleading pairs (same string, numbers and special characters) is reduced from 12.1% to 4.6%.

English	Italian	Latent Variable
dog	cane	Aligned
dog	cani	Aligned
dog	dog	Noise
good	buon	Aligned
good	buona	Aligned
good	santo	Noise
new	new	Noise
new	york	Noise
new	nuove	Aligned

Table 2: A sample of decisions from the noise-aware alignment on the English → Italian dataset.

5.2 Diachronic (Historical) Word Embedding

Experiment Setup The goal is to align English word-embedding derived from texts from different time periods, in order to identify which words changed meaning over time. The assumption is that most words remained stable, and hence the supervision is derived by aligning each word to itself. This problem contains noise in the lexicon by definition. We follow the exact setup fully described in Hamilton et al. (2016), but replace the OP algorithm with our Noise-aware version [1]. We project 1900s embeddings to 1990s embeddings vector-space. The top 10 distant word embeddings after alignment are analyzed by linguistic experts for semantic shift.

Experiment Results 45.5% of the input pairs were identified as noise. After the post processing of removing the non-frequent words as described in the experiment setup we end up with 121 noisy words. Our algorithm successfully identifies **all** the top-changing words in Hamilton et al. (2016) as noise, and learns to ignore them in the alignment. In addition, we argue our method provides better alignment. Table 3 shows the Nearest Neighbor (NN) of a 1990s word, in the 1900s vector-space after projection. We look at the top 10 changed words in Hamilton et al. (2016) and 3 unchanged words. We compare the alignment of the OP projection to the Noise-aware Alignment (NAA). For example, with our solution the word *actually* whose meaning shifted from "in fact" to express emphasize or surprise, is correctly mapped to *really* instead of *believed*. The word *gay* shifted from *cheerful* to *homosexual*, yet is still mapped to *gay* with NAA. This happens because the related embeddings (*homosexual*, *lesbian* and so on) are empty embeddings in 1900s, leaving *gay* as the next-best candidate, which we argue is better than OP's *society*. The words *car, driver, eve* whose meaning didn't change, were incorrectly aligned with OP to *cab, stepped, anniversary* instead of to themselves.

[1] Pre-possessing: removing proper nouns, stop words and empty embeddings. Post-processing: removing words whose frequency is below 10^{-5} in either years.

1990s Word	1900s NN aligned with OP	1900s NN aligned with NAA	Latent Variable
wanting	need	wishing	Noise
gay	society	gay	Noise
check	give	send	Noise
starting	begin	beginning	Noise
major	general	successful	Noise
actually	believed	really	Noise
touching	touched	touching	Noise
harry	hello	john	Noise
headed	halfway	toward	Noise
romance	artists	romance	Noise
car	cab	car	Aligned
driver	stepped	driver	Aligned
eve	anniversary	eve	Aligned

Table 3: Diachronic Semantic Change Experiment. Upper-part: noisy pairs. Bold: real semantic shifts. Underlined: global genre/discourse shifts. Unmarked: corpus artifacts. Bottom-part: clean pairs: Italics: unchanged words, no semantic shift.

6 Conclusion

We introduced the problem of embedding space projection with *noisy* lexicons, and showed that existing projection methods are sensitive in the presence of noise. We proposed an EM algorithm that jointly learns the projection and identifies the noisy pairs. The algorithm can be used as a drop-in replacement for the OP algorithm, and was demonstrated to improve results on two NLP tasks. We provide code at https://github.com/NoaKel/Noise-Aware-Alignment.

Acknowledgments

The work was supported by The Israeli Science Foundation (grant number 1555/15), and by the Israeli ministry of Science, Technology and Space through the Israeli-French Maimonide Cooperation program. We also, thank Roee Aharoni for helpful discussions and suggestions.

References

David Alvarez-Melis and Tommi S Jaakkola. 2018. Gromov-wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013*.

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017a. 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)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017b. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Projecting embeddings for domain adaption: Joint modeling of sentiment analysis in diverse domains. *arXiv preprint arXiv:1806.04381*.

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

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Georgiana Dinu and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *CoRR*, abs/1412.6568.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

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

Sebastian Ruder, Ryan Cotterell, Yova Kementchedjhieva, and Anders Søgaard. 2018. A discriminative latent-variable model for bilingual lexicon induction. *arXiv preprint arXiv:1808.09334*.

Peter Schnemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

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.

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)*, volume 1, pages 247–257.

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

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. *arXiv preprint arXiv:1809.03633*.

Josuke Yamane, Tomoya Takatani, Hitoshi Yamada, Makoto Miwa, and Yutaka Sasaki. 2016. Distributional hypernym generation by jointly learning clusters and projections. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1871–1879.

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)*, volume 1, pages 1959–1970.