Proceedings Paper:
Peng, X., Lin, C. and Stevenson, R. (2021) Cross-lingual word embedding refinement by ℓ1 norm optimisation. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T. and Zhou, Y., (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 06-11 Jun 2021, Virtual conference. The Association for Computational Linguistics. Article no: 214, pp. 2690-2701. ISBN: 9781954085466.

https://doi.org/10.18653/v1/2021.naacl-main.214

Cross-Lingual Word Embedding Refinement by ℓ_1 Norm Optimisation

Xutan Peng **Chenghua Lin*** **Mark Stevenson**
Department of Computer Science, The University of Sheffield, UK
{x.peng, c.lin, mark.stevenson}@shef.ac.uk

Abstract

Cross-Lingual Word Embeddings (CLWEs) encode words from two or more languages in a shared high-dimensional space in which vectors representing words with similar meaning (regardless of language) are closely located. Existing methods for building high-quality CLWEs learn mappings that minimise the ℓ_2 norm loss function. However, this optimisation objective has been demonstrated to be sensitive to outliers. Based on the more robust Manhattan norm (aka. ℓ_1 norm) goodness-of-fit criterion, this paper proposes a simple post-processing step to improve CLWEs. An advantage of this approach is that it is fully agnostic to the training process of the original CLWEs and can therefore be applied widely. Extensive experiments are performed involving ten diverse languages and embeddings trained on different corpora. Evaluation results based on bilingual lexicon induction and cross-lingual transfer for natural language inference tasks show that the ℓ_1 refinement substantially outperforms four state-of-the-art baselines in both supervised and unsupervised settings. It is therefore recommended that this strategy be adopted as a standard for CLWE methods.

1 Introduction

Cross-Lingual Word Embedding (CLWE) techniques have recently received significant attention as an effective means to support Natural Language Processing applications for low-resource languages, e.g., machine translation (Artetxe et al., 2018b) and transfer learning (Peng et al., 2021).

The most successful CLWE models are the so-called *projection-based* methods, which learn mappings between monolingual word vectors with very little, or even zero, cross-lingual supervision (Lample et al., 2018; Artetxe et al., 2018a; Glavaš et al., 2019). Mainstream projection-based CLWE models typically identify orthogonal mappings by minimising the topological dissimilarity between source and target embeddings based on ℓ_2 loss (aka. Frobenius loss or squared error) (Glavaš et al., 2019; Ruder et al., 2019). This learning strategy has two advantages. First, adding the orthogonality constraint to the mapping function has been demonstrated to significantly enhance the quality of CLWEs (Xing et al., 2015). Second, the existence of a closed-form solution to the ℓ_2 optima (Schönemann, 1966) greatly simplifies the computation required (Artetxe et al., 2016; Smith et al., 2017).

Despite its popularity, work in various application domains has noted that ℓ_2 loss is not robust to noise and outliers. It is widely known in computer vision that ℓ_2-loss-based solutions can severely exaggerate noise, leading to inaccurate estimates (Aanæs et al., 2002; De La Torre and Black, 2003). In data mining, Principal Component Analysis (PCA) using ℓ_2 loss has been shown to be sensitive to the presence of outliers in the input data, degrading the quality of the feature space produced (Kwak, 2008). Previous studies have demonstrated that the processes used to construct monolingual and cross-lingual embeddings may introduce noise (e.g. via reconstruction error (Allen and Hospedales, 2019) and structural variance (Ruder et al., 2019)), making the presence of outliers more likely. Empirical analysis of CLWEs also demonstrates that more distant word pairs (which are more likely to be outliers) have more influence on the behaviour of ℓ_2 loss than closer pairs. This raises the question of the appropriateness of ℓ_2 loss functions for CLWEs.

Compared to the conventional ℓ_2 loss, ℓ_1 loss (aka. Manhattan distance) has been mathematically demonstrated to be less affected by outliers (Rousseeuw and Leroy, 1987) and empirically proven useful in computer vision and data mining (Aanæs et al., 2002; De La Torre and Black, 2003; Kwak, 2008). Motivated by this insight, our paper proposes a simple yet effective post-

*Chenghua Lin is the corresponding author.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2690–2701
June 6–11, 2021.

processing technique to improve the quality of CLWEs: adjust the alignment of *any* cross-lingual vector space to minimise the ℓ_1 loss without violating the orthogonality constraint. Specifically, given existing CLWEs, we bidirectionally retrieve bilingual vectors and optimise their Manhattan distance using a numerical solver. The approach can be applied to any CLWEs, making the post-hoc refinement technique generic and applicable to a wide range of scenarios. We believe this to be the first application of ℓ_1 loss to the CLWE problem.

To demonstrate the effectiveness of our method, we select four state-of-the-art baselines and conduct comprehensive evaluations in both supervised and unsupervised settings. Our experiments involve ten languages from diverse branches/families and embeddings trained on corpora of different domains. In addition to the standard Bilingual Lexicon Induction (BLI) benchmark, we also investigate a downstream task, namely cross-lingual transfer for Natural Language Inference (NLI). In all setups tested, our algorithm significantly improves the performance of strong baselines. Finally, we provide an intuitive visualisation illustrating why ℓ_1 loss is more robust than it ℓ_2 counterpart when refining CLWEs (see Fig. 1). Our code is available at `https://github.com/Pzoom522/L1-Refinement`.

Our contribution is three-fold: (1) we propose a robust refinement technique based on the ℓ_1 norm training objective, which can effectively enhance CLWEs; (2) our approach is generic and can be directly coupled with both supervised and unsupervised CLWE models; (3) our ℓ_1 refinement algorithm achieves state-of-the-art performance for both BLI and cross-lingual transfer for NLI tasks.

2 Related Work

CLWE methods. One approach to generating CLWEs is to train shared semantic representations using multilingual texts aligned at sentence or document level (Vulić and Korhonen, 2016; Upadhyay et al., 2016). Although this research direction has been well studied, the parallel setup requirement for model training is expensive, and hence impractical for low-resource languages.

Recent years have seen an increase in interest in projection-based methods, which train CLWEs by finding mappings between pretrained word vectors of different languages (Mikolov et al., 2013; Lample et al., 2018; Peng et al., 2020). Since the input embeddings can be generated independently using monolingual corpora only, projection-based methods reduce the supervision required for training and offer a viable solution for low-resource scenarios.

Xing et al. (2015) showed that the precision of the learned CLWEs can be improved by constraining the mapping function to be orthogonal, which is formalised as the so-called ℓ_2 Orthogonal Procrustes Analysis (OPA):

$$\underset{\mathbf{M}\in\mathcal{O}}{\operatorname{argmin}} \|\mathbf{AM}-\mathbf{B}\|_2, \quad (1)$$

where $\mathbf{M}$ is the CLWE mapping, $\mathcal{O}$ denotes the orthogonal manifold (aka. the Stiefel manifold (Chu and Trendafilov, 2001)), and $\mathbf{A}$ and $\mathbf{B}$ are matrices composed using vectors from source and target embedding spaces.

While Xing et al. (2015) exploited an approximate and relatively slow gradient-based solver, more recent approaches such as Artetxe et al. (2016) and Smith et al. (2017) introduced an exact closed-form solution for Eq. (1). Originally proposed by Schönemann (1966), it utilises Singular Value Decomposition (SVD):

$$\mathbf{M}^\star = \mathbf{UV}^\intercal, \text{with } \mathbf{U}\Sigma\mathbf{V}^\intercal = \text{SVD}(\mathbf{A}^\intercal\mathbf{B}), \quad (2)$$

where $\mathbf{M}^\star$ denotes the ℓ_2-optimal mapping matrix. The efficiency and effectiveness of Eq. (2) have led to its application within many other approaches, e.g., Ruder et al. (2018), Joulin et al. (2018) and Glavaš et al. (2019). In particular, PROC-B (Glavaš et al., 2019), a supervised CLWE framework that simply applies multiple iterations of ℓ_2 OPA, has been demonstrated to produce very competitive performance on various benchmark tasks including BLI as well as cross-lingual transfer for NLI and information retrieval.

While the aforementioned approaches still require some weak supervision (i.e., seed dictionaries), there have also been some successful attempts to train CLWEs in a completely unsupervised fashion. For instance, Lample et al. (2018) proposed a system called MUSE, which bootstraps CLWEs without any bilingual signal through adversarial learning. VECMAP (Artetxe et al., 2018a) applied a self-learning strategy to iteratively compute the optimal mapping and then retrieve bilingual dictionary. Comparing MUSE and VECMAP, the latter tends to be more robust as its similarity-matrix-based heuristic initialisation is more stable in most cases (Glavaš et al., 2019; Ruder et al., 2019). Very

recently, some studies bootstrapped unsupervised CLWEs by jointly training word embeddings on concatenated corpora of different languages and achieved good performance (Wang et al., 2020).

The ℓ_2 refinement algorithm. CLWE models often apply ℓ_2 refinement, a post-processing step shown to improve the quality of the initial alignment (see Ruder et al. (2019) for survey). Given existing CLWEs $\{\mathbf{X}_{\mathtt{L_A}}, \mathbf{X}_{\mathtt{L_B}}\}$ for languages $\mathtt{L_A}$ and $\mathtt{L_B}$, bidirectionally one can use approaches such as the classic nearest-neighbour algorithm, the inverted softmax (Smith et al., 2017) and the cross-domain similarity local scaling (CSLS) (Lample et al., 2018) to retrieve two bilingual dictionaries $D_{\mathtt{L_A}\mapsto\mathtt{L_B}}$ and $D_{\mathtt{L_B}\mapsto\mathtt{L_A}}$. Note that word pairs in $D_{\mathtt{L_A}\mapsto\mathtt{L_B}} \cap D_{\mathtt{L_B}\mapsto\mathtt{L_A}}$ are highly reliable, as they form "mutual translations". Next, one can compose bilingual embedding matrices $\mathbf{A}$ and $\mathbf{B}$ by aligning word vectors (rows) using the above word pairs. Finally, a new orthogonal mapping is learned to fit $\mathbf{A}$ and $\mathbf{B}$ based on least-square regressions, i.e., perform ℓ_2 OPA described in Eq. (1).

Early applications of ℓ_2 refinement applied a *single* iteration, e.g. (Vulić and Korhonen, 2016). Due to the wide adoption of the closed-form ℓ_2 OPA solution (cf. Eq. (2)), recent methods perform multiple iterations. The iterative ℓ_2 refinement strategy is an important component of approaches that bootstrap from small or null training lexicons (Artetxe et al., 2018a). However, a single step of refinement is often sufficient to create suitable CLWEs (Lample et al., 2018; Glavaš et al., 2019).

3 Methodology

A common characteristic of CLWE methods that apply the orthogonality constraint is that they optimise using ℓ_2 loss (see § 2). However, outliers have disproportionate influence in ℓ_2 since the penalty increases quadratically and this can be particularly problematic with noisy data since the solution can "shift" towards them (Rousseeuw and Leroy, 1987). The noise and outliers present in real-world word embeddings may affect the performance of ℓ_2-loss-based CLWEs.

The ℓ_1 norm cost function is more robust than ℓ_2 loss as it is less affected by outliers (Rousseeuw and Leroy, 1987). Therefore, we propose a refinement algorithm for improving the quality of CLWEs based on ℓ_1 loss. This novel method, which we refer to as ℓ_1 refinement, is generic and can be applied post-hoc to improve the output of existing CLWE models. To our knowledge, the use of alternatives to ℓ_2-loss-based optimisation has never been explored by the CLWE community.

To begin with, analogous to ℓ_2 OPA (cf. Eq. (1)), ℓ_1 OPA can be formally defined and rewritten as

$$\begin{aligned}&\operatorname*{argmin}_{\mathbf{M}\in\mathcal{O}} \|\mathbf{AM}-\mathbf{B}\|_1 \\ &= \operatorname*{argmin}_{\mathbf{M}\in\mathcal{O}} \operatorname{tr}[(\mathbf{AM}-\mathbf{B})^\intercal \operatorname{sgn}(\mathbf{AM}-\mathbf{B})], \quad (3)\end{aligned}$$

where $\operatorname{tr}(\cdot)$ returns the matrix trace, $\operatorname{sgn}(\cdot)$ is the signum function, and $\in \mathcal{O}$ denotes that M is subject to the orthogonal constraint. Compared to ℓ_2 OPA which has a closed-form solution, solving Eq. (3) is much more challenging due to the discontinuity of $\operatorname{sgn}(\cdot)$. This issue can be addressed by replacing $\operatorname{sgn}(\cdot)$ with $\tanh(\alpha(\cdot))$, a smoothing function parameterised by α, such that

$$\operatorname*{argmin}_{\mathbf{M}\in\mathcal{O}} \operatorname{tr}[(\mathbf{AM}-\mathbf{B})^\intercal \tanh(\alpha(\mathbf{AM}-\mathbf{B}))]. \quad (4)$$

Larger values for α lead to closer approximations to $\operatorname{sgn}(\cdot)$ but reduce the smoothing effect. This approach has been used in many applications, such as the activation function of long short-term memory networks (Hochreiter and Schmidhuber, 1997).

However, in practice, we find that Eq. (4) remains unsolvable in our case with standard gradient-based frameworks for two reasons. First, α has to be sufficiently large in order to achieve a good approximation of $\operatorname{sgn}(\cdot)$. Otherwise, relatively small residuals will be down-weighted during fitting and the objective will become biased towards outliers, just similar to ℓ_2 loss. However, satisfying this requirement (i.e., large α) will lead to the activation function $\tanh(\alpha(\cdot))$ becoming easily *saturated*, resulting in an optimisation process that becomes trapped during the early stages. In other words, the optimisation can only reach an unsatisfactory local optimum. Second, the orthogonality constraint (i.e., $\mathbf{M} \in \mathcal{O}$) also makes the optimisation more problematic for these methods.

We address these challenges by adopting the approaches proposed by Trendafilov (2003). This method explicitly encourages the solver to only explore the desired manifold $\mathcal{O}$ thereby reducing the ℓ_1 solver's search space and difficulty of the optimisation problem. We begin by calculating the gradient ∇ w.r.t. the objective in Eq. (4) through matrix differentiation:

$$\nabla = \mathbf{A}^\intercal(\tanh(\mathbf{Z}) + \mathbf{Z} \odot \cosh^{-2}(\mathbf{Z})), \quad (5)$$

where $\mathbf{Z}{=}\alpha(\mathbf{AM}{-}\mathbf{B})$ and $\odot$ is the Hadamard product. Next, to find the steepest descent direction while ensuring that any $\mathbf{M}$ produced is orthogonal, we project ∇ onto $\mathcal{O}$, yielding[1]

$$\pi_{\mathcal{O}}(\nabla){:=}\frac{1}{2}\mathbf{M}(\mathbf{M}^{\intercal}\nabla{-}\nabla^{\intercal}\mathbf{M}){+}(\mathbf{I}{-}\mathbf{M}\mathbf{M}^{\intercal})\nabla. \quad (6)$$

Here $\mathbf{I}$ is an identity matrix with the shape of $\mathbf{M}$. With Eq. (6) defining the optimisation flow, our ℓ_1 loss minimisation problem reduces to an integration problem, as

$$\mathbf{M}^{\star} = \mathbf{M_0} + \int - \pi_{\mathcal{O}}(\nabla)\,\mathrm{d}t, \quad (7)$$

where $\mathbf{M_0}$ is a proper initial solution of Eq. (3) (e.g., ℓ_2-optimal mapping obtained via Eq. (2)).

Empirically, unlike the aforementioned standard gradient-based methods, by following the established policy of Eq. (6), the optimisation process of Eq. (7) will not violate the orthogonality restriction or get trapped during early stages. However, this ℓ_1 OPA solver requires extremely small step size to generate reliable solutions (Trendafilov, 2003), making it computationally expensive[2]. Therefore, it is impractical to perform ℓ_1 refinement in an iterative fashion like ℓ_2 refinement without significant computational resources.

Previous work has demonstrated that applying the ℓ_1-loss-based algorithms from a good initial state can speed up the optimisation. For instance, Kwak (2008) found that feature spaces created by ℓ_2 PCA were severely affected by noise. Replacing the cost function with ℓ_1 loss significantly reduced this problem, but required expensive linear programming. To reduce the convergence time, Brooks and Jot (2013) exploited the first principal component from the ℓ_2 solution as an initial guess. Similarly, when reconstructing corrupted pixel matrices, ℓ_2-loss-based results are far from satisfactory; using ℓ_1 norm estimators can improve the quality, but are too slow to handle large-scale datasets (Aanæs et al., 2002). However, taking the ℓ_2 optima as the starting point allowed less biased reconstructions to be learned in an acceptable time (De La Torre and Black, 2003).

Inspired by these works, we make use of ℓ_1 refinement to carry out post-hoc enhancement of existing CLWEs. Our full pipeline is described in

[1] See Chu and Trendafilov (2001) for derivation details.

[2] It takes averagely 3 hours and up to 12 hours to perform Eq. (7) on an Intel Core i9-9900K CPU. In comparison, the time required to solve Eq. (2) in each training loop is less than 1 second and the iterative ℓ_2-norm-based training takes 1 to 5 hours in total.

Algorithm 1 ℓ_1 refinement

Input: CLWEs $\{\mathbf{X}_{\mathtt{L_A}}, \mathbf{X}_{\mathtt{L_B}}\}$
Output: updated CLWEs $\{\mathbf{X}_{\mathtt{L_A}}\mathbf{M}^{\star}, \mathbf{X}_{\mathtt{L_B}}\}$
1: $D_{\mathtt{L_A}\mapsto\mathtt{L_B}} \leftarrow$ build dict via $\mathbf{X}_{\mathtt{L_A}}$ and $\mathbf{X}_{\mathtt{L_B}}$
2: $D_{\mathtt{L_B}\mapsto\mathtt{L_A}} \leftarrow$ build dict via $\mathbf{X}_{\mathtt{L_B}}$ and $\mathbf{X}_{\mathtt{L_A}}$
3: $D \leftarrow D_{\mathtt{L_A}\mapsto\mathtt{L_B}} \cap D_{\mathtt{L_B}\mapsto\mathtt{L_A}}$
4: $\mathbf{A}, \mathbf{B} \leftarrow$ looks up for D in $\mathbf{X}_{\mathtt{L_A}}, \mathbf{X}_{\mathtt{L_B}}$
5: perform integration to solve Eq. (7) for $\mathbf{M}^{\star}$, with initial value $\mathbf{M_0} \leftarrow \mathbf{I}$, until stopping criteria are met

Algorithm 1 (see § 4.3 for implemented configurations). In common with ℓ_2 refinement (cf. § 2), steps 1-4 bootstrap a synthetic dictionary D and compose bilingual word vector matrices $\mathbf{A}$ and $\mathbf{B}$ which have reliable row-wise correspondence. Taking them as the starting state, in step 5 an identity matrix naturally serves as our initial solution $\mathbf{M_0}$.

During the execution of Eq. (7), we record ℓ_1 loss per iteration and see if *either* of the following two stopping criteria have been satisfied: (1) the updated ℓ_1 loss exceeds that of the previous iteration; (2) on-the-fly $\mathbf{M}$ has non-negligibly departed from the orthogonal manifold, which can be indicated by the maximum value of the disparity matrix as

$$\max(|\mathbf{M}^{\intercal}\mathbf{M} - \mathbf{I}|) > \epsilon, \quad (8)$$

where ϵ is a sufficiently small threshold. The resulting $\mathbf{M}^{\star}$ can be used to adjust the word vectors of $\mathtt{L_A}$ and output refined CLWEs.

A significant advantage of our algorithm is its generality: it is fully independent of the method used for creating the original CLWEs and can therefore be used to enhance a wide range of models, both in supervised and unsupervised settings.

4 Experimental Setup

4.1 Datasets

In order to demonstrate the generality of our proposed method, we conduct experiments using two groups of monolingual word embeddings trained on very different corpora:

Wiki-Embs (Grave et al., 2018): embeddings developed using Wikipedia dumps for a range of ten diverse languages: two Germanic (English|EN, German|DE), two Slavic (Croatian|HR, Russian|RU), three Romance (French|FR, Italian|IT, Spanish|ES) and three non-Indo-European (Finnish|FI from the Uralic family, Turkish|TR from the Turkic family and Chinese|ZH from the Sino-Tibetan family).

News-Embs (Artetxe et al., 2018a): embeddings trained on a multilingual News text collection, i.e.,

the WaCKy Crawl of {EN, DE, IT}, the Common Crawl of FI, and the WMT News Crawl of ES.

News-Embs are considered to be more challenging for building good quality CLWEs due to the heterogeneous nature of the data, while a considerable portion of the multilingual training corpora for Wiki-Embs are roughly parallel. Following previous studies (Lample et al., 2018; Artetxe et al., 2018a; Zhou et al., 2019; Glavaš et al., 2019), only the first 200K vocabulary entries are preserved.

4.2 Baselines

Glavaš et al. (2019) provided a systematic evaluation for projection-based CLWE models, demonstrating that three methods (i.e., MUSE, VECMAP, and PROC-B) achieve the most competitive performance. A recent algorithm (JA) by Wang et al. (2020) also reported state-of-the-art results. For comprehensive comparison, we therefore use all these four methods as the main baselines for both supervised and unsupervised settings:

MUSE (Lample et al., 2018): an *unsupervised* CLWE model based on adversarial learning and iterative ℓ_2 refinement;

VECMAP (Artetxe et al., 2018a): a robust *unsupervised* framework using a self-learning strategy;

PROC-B (Glavaš et al., 2019): a simple but effective *supervised* approach to creating CLWEs;

JA-MUSE and **JA-RCSLS** (Wang et al., 2020): a recently proposed Joint-Align (**JA**) Framework, which first initialises CLWEs using joint embedding training, followed by vocabularies reallocation. It then utilises off-the-shelf CLWE methods to improve the alignment in both *unsupervised* (**JA-MUSE**) and *supervised* (**JA-RCSLS**) settings.

In the original implementations, MUSE, PROC-B and JA were only trained on Wiki-Embs while VECMAP additionally used News-Embs. Although all baselines reported performance for BLI, they used various versions of evaluation sets, hence previous results are not directly comparable with the ones reposted here. More concretely, the testsets for MUSE/JA and VECMAP are two different batches of EN-centric dictionaries, while the testset for PROC-B also supports non-EN translations.

4.3 Implementation Details of Algorithm 1

The CSLS scheme with a neighbourhood size of 10 (CSLS-10) is adopted to build synthetic dictionaries via the input CLWEs. A variable-coefficient ordinary differential equation (VODE) solver[3] was implemented for the system described in Eq. (7). Suggested by Trendafilov (2003), we set the maximum order at 15, the smoothness coefficient α in Eq. (5) at 1e8, the threshold ϵ in Eq. (8) at 1e-5, and performed the integration with a fixed time interval of 1e-6. An early-stopping design was adopted to ensure computation completed in a reasonable time: in addition to the two default stopping criteria in § 3, integration is terminated if $\int \mathrm{d}t$ reaches 5e-3 ($\mathrm{d}t$ is the differentiation term in Eq. (7)).

In terms of the tolerance of the VODE solver, we set the absolute tolerance at 1e-7 and the relative tolerance at 1e-5, following the established approach of Kulikov (2013). These tolerance settings show good generality empirically and were used for all tested language pairs, datasets, and models in our experiments.

5 Results

We evaluate the effectiveness of the proposed ℓ_1 refinement technique on two benchmarks: Bilingual Lexicon Induction (BLI), the *de facto* standard for measuring the quality of CLWEs, and a downstream natural language inference task based on cross-lingual transfer. In addition to comparison against state-of-the-art CLWE models, we also report the performance of the single-iteration ℓ_2 refinement method which follows steps 1-4 of Algorithm 1 then minimises ℓ_2 loss in the final step.

To reduce randomness, we executed each model in each setup three times and the average accuracy (ACC, aka. precision at rank 1) is reported. Following Glavaš et al. (2019), by comparing scores achieved before and after ℓ_1 refinement, statistical significance is indicated via the p-value of two-tailed t-tests with Bonferroni correction (Dror et al., 2018) (note that p-values are not recorded for Tab. 2b given the small number of runs).

5.1 Bilingual Lexicon Induction

Refining unsupervised baselines. Tab. 1a follows the main setup of Lample et al. (2018), who tested six language pairs using Wiki-Embs[4]. After ℓ_1 refinement, MUSE-ℓ_1, JA-MUSE-ℓ_1, and VECMAP-ℓ_1 all significantly ($p < 0.01$) outperform their corresponding base algorithms, with an average 1.1% performance gain over MUSE,

[3] http://www.netlib.org/ode/vode.f

[4] Note that we are unable to report the result of English to Esperanto as the corresponding dictionary is missing, see https://git.io/en-eo-dict-issue.

	EN–DE	EN–ES	EN–FR	EN–RU	EN–ZH
MUSE ♥	74.0	81.7	82.3	44.0	32.5
MUSE-ℓ_2	74.0	82.1	82.6	43.8*	31.9*
MUSE-ℓ_1	75.2	82.6	82.9	45.6*	33.8*
JA-MUSE ♦	74.2	81.4	82.8	45.0	36.1
JA-MUSE-ℓ_2	74.1	81.6	82.7	45.1	36.2
JA-MUSE-ℓ_1	**75.4**	82.0	**83.1**	46.3	**38.1**
VECMAP ♣	75.1	82.3	80.0	49.2	00.0
VECMAP-ℓ_2	74.8	82.3	79.4	48.9	00.0
VECMAP-ℓ_1	**75.4**	**82.9**	80.2	**49.9**	00.0

(a) Wiki-Embs (setup of Lample et al. (2018)).

	EN–DE	EN–ES	EN–FI	EN–IT
MUSE ♥	00.0	07.1	00.0	09.1
MUSE-ℓ_2	00.0	00.0	00.0	00.0
MUSE-ℓ_1	00.0	00.0	00.0	00.0
JA-MUSE	47.9	48.4	33.0	37.2
JA-MUSE-ℓ_2	47.9	48.6	32.9	37.3
JA-MUSE-ℓ_1	48.8	**49.7**	**35.2**	37.7
VECMAP ♥	48.2	48.1	32.6	37.3
VECMAP-ℓ_2	48.1	47.9	32.9	37.1
VECMAP-ℓ_1	**49.0**	48.9	34.4	**37.8**

(b) News-Embs (setup of Artetxe et al. (2018a)).

Table 1: ACC (%) of unsupervised BLI. (a) Rows marked with ♥, ♦ and ♣ are respectively from Lample et al. (2018), Wang et al. (2020) and Zhou et al. (2019). NB: for EN–{RU, ZH} we observed one failed run (ACC <10.0%), where we only record the average of successful scores with *. (b) Rows marked with ♥ are from Artetxe et al. (2018a).

	EN–DE	EN–FI	EN–FR	EN–HR	EN–IT	EN–RU	EN–TR
JA-RCSLS	50.9	33.9	63.0	29.1	58.3	41.3	29.4
JA-RCSLS-ℓ_2	50.7	33.8	63.0	29.1	58.2	41.3	29.5
JA-RCSLS-ℓ_1	51.6	34.5	63.4	30.4	59.0	41.9	30.2
PROC-B ♥	52.1	36.0	63.3	29.6	**60.5**	41.9	30.1
PROC-B-ℓ_2	51.8	34.4	63.1	28.2	**60.5**	39.8	28.0
PROC-B-ℓ_1	**52.6**	**36.3**	**63.7**	**30.5**	**60.5**	**42.3**	**30.9**

(a) Wiki-Embs (setup of Glavaš et al. (2019)).

	EN–DE	EN–FI	EN–IT
JA-RCSLS	46.8	42.0	37.4
JA-RCSLS-ℓ_2	46.9	42.2	37.5
JA-RCSLS-ℓ_1	48.3	**44.6**	39.0
PROC-B	47.5	41.4	37.3
PROC-B-ℓ_2	47.1	41.7	37.4
PROC-B-ℓ_1	**52.6**	43.3	**41.1**

(b) News-Embs.

Table 2: MRR (%) of supervised BLI. Rows marked with ♥ are from the supplementary of Glavaš et al. (2019).

1.1% over JA-MUSE, and 0.5% over VECMAP. To put these improvements in context, Heyman et al. (2019) reported an improvement of 0.4% for VECMAP on same dataset and language pairs.

Our method tends to work better on the more distant language pairs. For instance, for the distant pairs EN–{RU, ZH}, the increments achieved by MUSE-ℓ_1 are 1.6% and 1.3%, respectively; whereas for the close pairs EN–{DE, ES, FR} the average gain is a maximum of 0.9%. A similar trend can be observed for JA-MUSE-ℓ_1 and VECMAP-ℓ_1. (As the VECMAP algorithm always collapses for EN–ZH, no result is reported for this language pair).

Another set of experiments were conducted to evaluate the robustness of our algorithm following the main setup of Artetxe et al. (2018a), who tested four language pairs based on the more homogeneous News-Embs. Tab. 1b shows that JA-MUSE-ℓ_1 and VECMAP-ℓ_1 consistently improves the original VECMAP with an average gain of 1.2% and 1.0% (p<0.01). Obtaining such substantial improvements over the state-of-the-art is nontrivial. For example, even a very recent weakly supervised method by Wang et al. (2019) is *inferior* to VECMAP by 1.0% average ACC. On the other hand, MUSE fails to produce any analysable result as it always collapses on the more challenging News-Embs. Improvement with ℓ_1 refinement is also larger when language pairs are more distant, e.g., for VECMAP-ℓ_1 the ACC gain on EN-FI is 1.8%, more than double of the gain (0.7%) on the close pairs EN–{DE, IT} (cf. Tab. 1a and above).

We also conduct an ablation study by reporting the performance of ℓ_2 refinement scheme ({MUSE, JAMUSE, VECMAP}-ℓ_2). This observation is in accordance with that of Lample et al. (2018), who reported that after performing ℓ_2 refinement in the first loop, applying further iterations only produces marginal precision gain, if any.

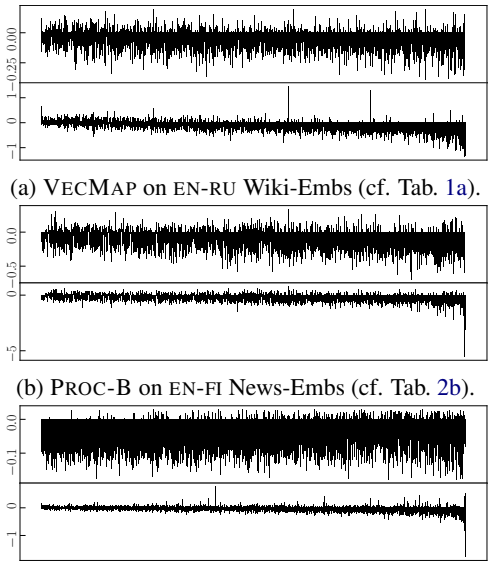
Overall, the ℓ_1 refinement consistently and significantly improve the CLWEs produced by base algorithms, regardless of the embeddings and setups used, thereby demonstrating the effectiveness and robustness of the proposed algorithm.

Refining supervised baselines. To test the generalisability of our method, we also applied it on state-of-the-art supervised CLWE models: PROC-B (Glavaš et al., 2019) and JA-RCSLS (Wang et al., 2020). Following the setup of Glavaš et al. (2019), we learn mappings using Wiki-Embs and 1K training splits of their dataset.

Their evaluation code retrieves bilingual word pairs using the classic nearest-neighbour algorithm and outputs the Mean Reciprocal Rank (MRR). As shown in Tab. 2a, both JA-RCSLS-ℓ_1 and PROC-B-ℓ_1 outperform the baseline algorithms for all

Unsupervised	DE–IT	DE–TR	FI–HR	FI–IT	HR–RU	IT–FR	TR–IT
ICP♥	44.7	21.5	20.8	26.3	30.9	62.9	24.3
GWA♥	44.0	10.1	00.9	17.3	00.1	65.5	14.2
MUSE♥	49.6	23.7	22.8	32.7	00.0	66.2	30.6
MUSE-ℓ_2	50.3	23.9	23.1	32.7	34.9	67.1	30.5*
MUSE-ℓ_1	50.7	26.5	25.4	35.0	37.9	67.6	33.3*
JA-MUSE	50.9	25.6	23.4	34.9	36.9	68.3	34.7
JA-MUSE-ℓ_2	50.9	25.5	23.4	34.7	36.9	68.4	34.7
JA-MUSE-ℓ_1	**51.5**	**28.4**	26.1	36.0	37.6	**68.7**	**36.1**
VECMAP♥	49.3	25.3	28.0	35.5	37.6	66.7	33.2
VECMAP-ℓ_2	48.8	25.7	28.5	35.8	38.4	67.0	33.5
VECMAP-ℓ_1	50.1	28.2	**30.3**	**37.1**	**40.1**	67.6	35.9
Supervised							
DLV♥	42.0	16.7	18.4	24.4	26.4	58.5	20.9
RCSLS♥	45.3	20.1	21.4	27.2	29.1	63.7	24.6
JA-RSCLS	46.6	20.9	22.1	29.0	29.9	65.2	25.3
JA-RSCLS-ℓ_2	46.4	20.8	22.3	29.0	29.8	65.2	25.3
JA-RSCLS-ℓ_1	47.3	22.2	23.8	30.1	31.2	65.9	26.6
PROC-B♥	50.7	25.0	26.3	32.8	34.8	66.5	29.8
PROC-B-ℓ_2	50.0	24.1	25.6	31.8	34.3	66.4	29.6
PROC-B-ℓ_1	**51.1**	**25.6**	**26.9**	**33.6**	**35.0**	**67.4**	**30.5**

Table 3: MRR (%) of BLI for non-EN language pairs. Rows marked with ♥ are from the supplementary of Glavaš et al. (2019). MUSE yielded one unsuccessful run for TR–IT, and we only record the average of the two successful scores with *.



(a) VECMAP on EN-RU Wiki-Embs (cf. Tab. 1a).

(b) PROC-B on EN-FI News-Embs (cf. Tab. 2b).

(c) MUSE on IT-FR Wiki-Embs (cf. Tab. 3).

Figure 1: Changes to $||\mathbf{AM} - \mathbf{B}||_2$ after applying ℓ_1 (upper) and ℓ_2 (lower) refinement. Each word pairs is represented by a bar ordered on the x-axis by the distance between them. See Fig. A.1 for the alternative version.

language pairs (with the exception of EN–IT where the score of PROC-B is unchanged) with an average improvement of 0.9% and 0.5%, respectively (p<0.01).

JA-RCSLS-ℓ_1 and PROC-B-ℓ_1 were also tested using News-Embs with results shown in Tab. 2b[5]. ℓ_1 refinement achieves an impressive improvement for both close (EN–{DE, IT}) and distant (EN–FI) language pairs: average gain of 1.9% and 3.9% respectively and over 5% for EN–DE (PROC-B-ℓ_1) in particular. The ℓ_2 refinement does not benefit the supervised baseline, similar to the lack of improvement observed in the unsupervised setups.

Comparison of unsupervised and supervised settings. This part provides a comparison of the effectiveness of ℓ_1 refinement in unsupervised and supervised scenarios. Unlike previous experiments where only alignments involving English were investigated, these tests focus on non-EN setups. Glavaš et al. (2019)'s dataset is used to construct seven representative pairs which cover every category of etymological combination, i.e., intra-language-branch {HR–RU, IT–FR}, inter-language-branch {DE–IT}, and inter-language-family {DE–TR, FI–HR, FI–IT, TR–IT}. The 1K training splits are used as seed lexicons in supervised runs. Apart from our main baselines, we further report the results of several other competitive CLWE models: Iterative Closest Point Model (ICP, Hoshen and Wolf, 2018), Gromov-Wasserstein Alignment Model (GWA, Alvarez-Melis and Jaakkola, 2018), Discriminative Latent-Variable Model (DLV, Ruder et al., 2018) and Relaxed CSLS Model (RCSLS, Joulin et al., 2018).

Results shown in Tab. 3 demonstrate that the main baselines (MUSE, JA-MUSE, VECMAP, JA-RCSLS, and PROC-B) outperform these other models by a large margin. For all these main baselines, post applying ℓ_1 refinement improves the mapping quality for all language pairs ($p < 0.01$), with average improvements of 1.7%, 1.4%, 1.8%, 1.1%, and 0.8%, respectively. Consistent with findings in the previous experiments, ℓ_2 refinement does not enhance performance. Improvement with ℓ_1 refinement is higher when language pairs are more distant, e.g., for all inter-language-family pairs such as FI–HR and TR–IT, even the minimum improvement of MUSE-ℓ_1 over MUSE is 2.3%.

Comparing unsupervised and supervised approaches, it can be observed that MUSE, JA-MUSE and VECMAP achieve higher overall gain with ℓ_1 refinement than JA-RCSLS and PROC-B, where JA-MUSE-ℓ_1 and VECMAP-ℓ_1 give the best overall performance. One possible explanation to this

[5]Note that results for EN–ES is not included, as no EN–ES dictionary is provided in Glavaš et al. (2019)'s dataset.

phenomenon is that there is only a single source of possible noise in unsupervised models (i.e. the embedding topology) but for supervised methods noise can also be introduced via the seed lexicons. Consequently unsupervised approaches drive more benefit from ℓ_1 refinement, which reduces the influence of topological outliers in CLWEs.

Topological behaviours of ℓ_1 and ℓ_2 refinements. To validate our assumption that ℓ_2 refinement is more sensitive to outliers while its ℓ_1 counterpart is more robust, we analyse how each refinement strategy changes the distance between bilingual word vector pairs in the synthetic dictionary D (cf. Algorithm 1) constructed from trained CLWE models. Specifically, for each word vector pair we subtract its post-refinement distance from the original distance (i.e., without applying additional ℓ_1 or ℓ_2 refinement step). Fig. 1 shows visualisation examples for three algorithms and language pairs, where each bar represents one word pair. It can be observed that ℓ_1 refinement effectively reduces the distance for most word pairs, regardless of their original distance (i.e., indicated by bars with negative values in the figures). The conventional ℓ_2 refinement strategy, in contrast, exhibits very different behaviour and tends to be overly influenced by word pairs with large distance (i.e. by outliers). The reason for this is that the ℓ_2-norm penalty increases quadratically, causing the solution to put much more weight on optimising distant word pairs (i.e., word pairs on the right end of the X-axis show sharp distance decrements). This observation is in line with Rousseeuw and Leroy (1987) and explains why ℓ_1 loss performs substantially stronger than ℓ_2 loss in the refinement.

Case study. After aligning EN-RU embeddings with unsupervised MUSE, we measured the distance between vectors corresponding to the ground-truth dictionary of Lample et al. (2018) (cf. Fig. 1a). We then detected large outliers by finding vector pairs whose distance falls above $\text{Q3} + 1.5 \cdot (\text{Q3} - \text{Q1})$, where Q1 and Q3 respectively denote the lower and upper quartile based on the popular Inter-Quartile Range (Hoaglin et al., 1986). We found that many of the outliers correspond to polysemous entries, such as {state (2× noun meanings and 1× verb meaning), состояние (only means *status*)}, {type (2× nominal meanings and 1× verb meaning), тип (only means *kind*)}, and {film (5× noun meanings), фильм (only means *movie*)}. We then re-perform ℓ_2-based mapping after removing these vector pairs, observing that the accuracy jumps to 45.9% (cf. the original ℓ_2-norm alignment it is 43.8% and after ℓ_1 refinement it is 45.6%, cf. Tab. 1). This indicates that although all baselines already make use of preprocessing steps including vector normalization, outlier issues still exist and harms the ℓ_2 norm CLWEs. However, they can be alleviated by the proposed ℓ_1 refinement technique.

Unsupervised	EN–DE	EN–FR	EN–RU	EN–TR
ICP♥	58.0	51.0	57.2	40.0
GWA♥	42.7	38.3	37.6	35.9
MUSE♥	61.1	53.6	36.3	35.9
MUSE-ℓ_2	61.1	53.0	57.3*	48.9*
MUSE-ℓ_1	**63.5**	55.3	58.9*	52.3*
JA-MUSE	61.3	55.2	58.1	55.0
JA-MUSE-ℓ_2	61.2	55.2	57.6	55.1
JA-MUSE-ℓ_1	62.9	57.9	59.4	**57.5**
VECMAP♥	60.4	61.3	58.1	53.4
VECMAP-ℓ_2	60.3	60.6	57.7	53.5
VECMAP-ℓ_1	61.5	**63.7**	**60.1**	56.4
Supervised				
RCSLS♥	37.6	35.7	37.8	38.7
JA-RSCLS	50.2	48.9	51.0	51.7
JA-RSCLS-ℓ_2	50.4	48.6	50.9	51.5
JA-RSCLS-ℓ_1	51.3	50.1	53.2	52.6
PROC-B♥	61.3	54.3	59.3	56.8
PROC-B-ℓ_2	61.0	**54.8**	58.9	55.1
PROC-B-ℓ_1	**62.1**	**54.8**	**60.7**	**58.2**

Table 4: ACC (%) of NLI. Rows marked with ♥ are from Glavaš et al. (2019). MUSE yielded one unsuccessful run for EN–RU and EN–TR respectively, which we exclude when calculating the average (with *).

5.2 Natural Language Inference

Finally, we experimented with a downstream NLI task in which the aim is to determine whether a "hypothesis" is true (*entailment*), false (*contradiction*) or undetermined (*neutral*), given a "premise". Higher ACC indicates better encoding of semantics in the tested embeddings. The CLWEs used are those trained with Wiki-Embs for BLI. For MUSE, JA-MUSE and VECMAP, we also obtain CLWEs for EN–TR pair with the same configuration.

Following Glavaš et al. (2019), we first train the Enhanced Sequential Inference Model (Chen et al., 2017) based on the large-scale English MultiNLI corpus (Williams et al., 2018) using vectors of language L_A (EN) from an aligned bilingual embedding space (e.g., EN–DE). Next, we replace the L_A vectors with the vectors of language L_B (e.g., DE), and directly test the trained model on the language L_B portion of the XNLI corpus (Conneau et al., 2018).

Results in Tab. 4 show that the CLWEs refined

by our algorithm yield the highest ACC for all language pairs in both supervised and unsupervised settings. The ℓ_2 refinement, on the contrary, is not beneficial overall. Improvements in cross-lingual transfer for NLI exhibit similar trends to those in the BLI experiments, i.e. greater performance gain for unsupervised methods and more distant language pairs, consistent with previous observations (Glavaš et al., 2019). For instance, MUSE-ℓ_1 JA-MUSE-ℓ_1 and VECMAP-ℓ_1 outperform their baselines by at least 2% in ACC on average ($p < 0.01$), whereas the improvements of JA-RSCLS-ℓ_1 and PROC-B-ℓ_1 over their corresponding base methods are 2% and 2.1% respectively ($p < 0.01$). For both unsupervised and supervised methods, ℓ_1 refinement demonstrates stronger effect for more distant language pairs, e.g., MUSE-ℓ_1 surpasses MUSE by 1.2% for EN–FR, whereas a more impressive 2.7% gain is achieved for EN–TR.

In summary, in addition to improving BLI performance, our ℓ_1 refinement method also produces a significant improvement for a downsteam task (NLI), demonstrating its effectiveness in improving the CLWE quality.

6 Conclusion and Future Work

This paper proposes a generic post-processing technique to enhance CLWE performance based on optimising ℓ_1 loss. This algorithm is motivated by successful applications in other research fields (e.g. computer vision and data mining) which exploit the ℓ_1 norm cost function since it has been shown to be more robust to noisy data than the commonly-adopted ℓ_2 loss. The approach was evaluated using ten diverse languages and word embeddings from different domains on the popular BLI benchmark, as well as a downstream task of cross-lingual transfer for NLI. Results demonstrated that our algorithm can significantly improve the quality of CLWEs in both supervised and unsupervised setups. It is therefore recommended that this straightforward technique be applied to improve performance of CLWEs.

The convergence speed of the optimiser prevented us from performing ℓ_1 loss optimisation over multiple iterations. Future work will focus on improving the efficiency of our ℓ_1 OPA solver, as well as exploring the application of other robust loss functions within CLWE training strategies.

Ethics Statement

This work provides an effective post-hoc method to improve CLWEs, advancing the state-of-the-art in both supervised and unsupervised settings. Our comprehensive empirical studies demonstrate that the proposed algorithm can facilitate researches in machine translation, cross-lingual transfer learning, etc, which have deep societal impact of bridging cultural gaps across the world.

Besides, this paper introduces and solves an optimisation problem based on an under-explored robust cost function, namely ℓ_1 loss. We believe it could be of interest for the wider community as outlier is a long-standing issue in many artificial intelligence applications.

One caveat with our method, as is the case for all word-embedding-based systems, is that various biases may exist in vector spaces. We suggest this problem should always be looked at critically. In addition, our implemented solver can be computationally expensive, leading to increased electricity consumption and the associated negative environmental repercussions.

Acknowledgements

This work is supported by the award made by the UK Engineering and Physical Sciences Research Council (Grant number: EP/P011829/1) and Baidu, Inc. We would also like to express our sincerest gratitude to Guanyi Chen, Ruizhe Li, Xiao Li, Shun Wang, and the anonymous reviewers for their insightful and helpful comments.

References

Henrik Aanæs, Rune Fisker, Kalle Åström, and Jens Michael Carstensen. 2002. Robust factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1215–1225.

Carl Allen and Timothy M. Hospedales. 2019. Analogies explained: Towards understanding word embeddings. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 223–231. PMLR.

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

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

J. Paul Brooks and Sapan Jot. 2013. pcaL1 : An implementation in r of three methods for ℓ_1-norm principal component analysis. In *Optimization Online preprint*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Moody T Chu and Nickolay T Trendafilov. 2001. The orthogonally constrained regression revisited. *Journal of Computational and Graphical Statistics*, pages 746–771.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Fernando De La Torre and Michael J Black. 2003. A framework for robust subspace learning. *International Journal of Computer Vision*.

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)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

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

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

Geert Heyman, Bregt Verreet, Ivan Vulić, and Marie-Francine Moens. 2019. Learning unsupervised multilingual word embeddings with incremental multilingual hubs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1890–1902, Minneapolis, Minnesota. Association for Computational Linguistics.

David C Hoaglin, Boris Iglewicz, and John W Tukey. 1986. Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*.

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

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

Gennady Yu Kulikov. 2013. Cheap global error estimation in some Runge–Kutta pairs. *IMA Journal of Numerical Analysis*.

N. Kwak. 2008. Principal component analysis based on ℓ_1-norm maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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

Xutan Peng, Chenghua Lin, Mark Stevenson, and Chen Li. 2020. Revisiting the linearity in cross-lingual embedding mappings: from a perspective of word analogies.

Xutan Peng, Yi Zheng, Chenghua Lin, and Advaith Siddharthan. 2021. Summarising historical text in modern languages. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3123–3142, Online. Association for Computational Linguistics.

Peter J. Rousseeuw and Annick M. Leroy. 1987. *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc., USA.

Sebastian Ruder, Ryan Cotterell, Yova Kementchedjhieva, and Anders Søgaard. 2018. A discriminative latent-variable model for bilingual lexicon induction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 458–468, Brussels, Belgium. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65(1).

Peter Schönemann. 1966. A generalized solution of the Orthogonal Procrustes Problem. *Psychometrika*.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Nickolay T. Trendafilov. 2003. On the ℓ_1 Procrustes problem. *Future Generation Computer Systems*. Selected papers on Theoretical and Computational Aspects of Structural Dynamical Systems in Linear Algebra and Control.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1661–1670, Berlin, Germany. Association for Computational Linguistics.

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)*, pages 247–257, Berlin, Germany. Association for Computational Linguistics.

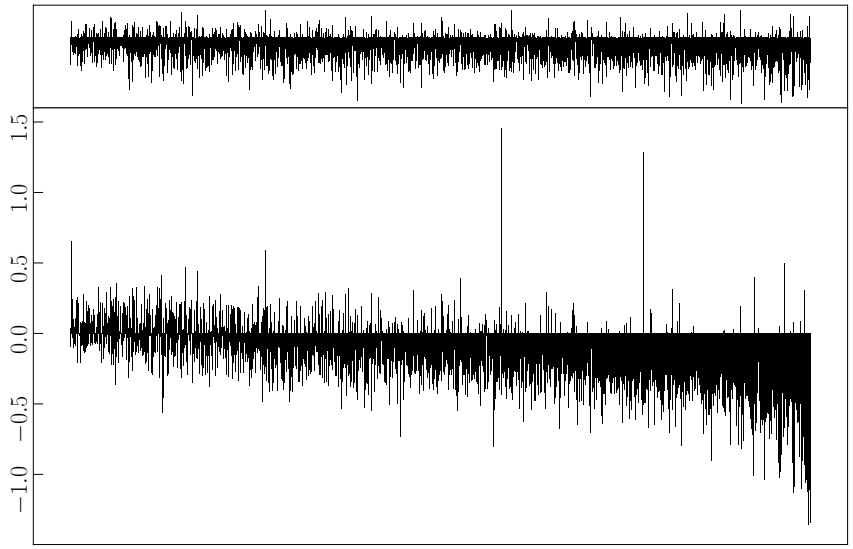
Haozhou Wang, James Henderson, and Paola Merlo. 2019. Weakly-supervised concept-based adversarial learning for cross-lingual word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4419–4430, Hong Kong, China. Association for Computational Linguistics.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and A simple unified framework. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

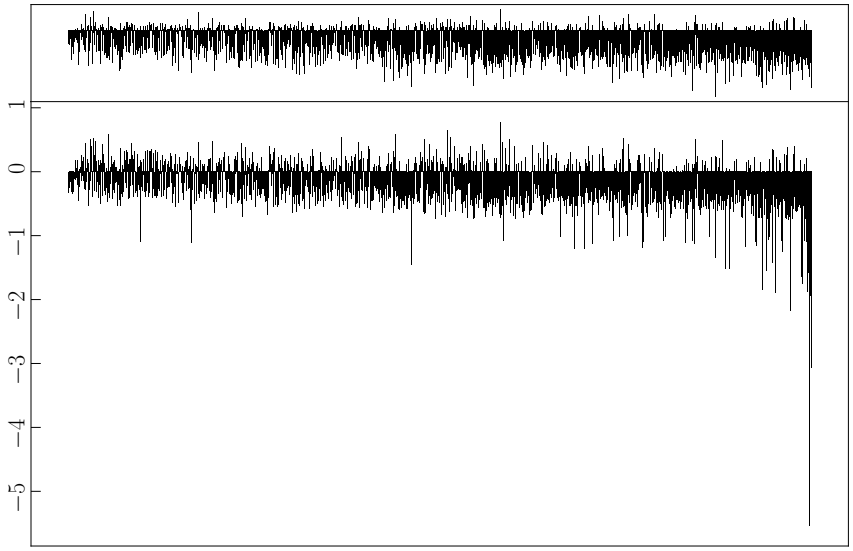
Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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

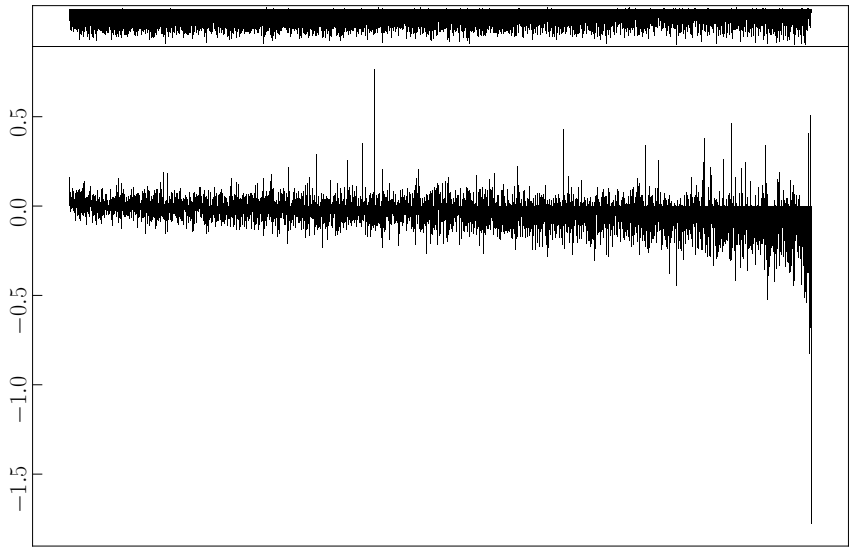
Chunting Zhou, Xuezhe Ma, Di Wang, and Graham Neubig. 2019. Density matching for bilingual word embedding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1588–1598, Minneapolis, Minnesota. Association for Computational Linguistics.



(a) VECMAP on EN-RU Wiki-Embs (cf. Tab. 1a).



(b) PROC-B on EN-FI News-Embs (cf. Tab. 2b).



(c) MUSE on IT-FR Wiki-Embs (cf. Tab. 3).

Figure A.1: Changes to $||\mathbf{AM} - \mathbf{B}||_2$ after applying ℓ_1 (upper) and ℓ_2 (lower) refinement. Different from Fig. 1, in each sub-figure the upper and lower Y-axis scales are uniformed.