Word Translation Without Parallel Data (2018)

A. Conneau, G. Lample, M. Ranzato, L. Denoyer, H. Jégou

Poster by: A. Puigdemont Monllor, J. Čieško

PA164: Machine Learning and NLP, Faculty of Informatics, Masaryk University, November 11, 2024

Objectives

- Propose an **unsupervised model** achieving state-of-the-art performance in cross-lingual tasks using monolingual corpora.
- Achieve word translation **without parallel data** by aligning monolingual embeddings with an **adversarial training method**.
- Improve performance through a **cross-domain similarity adjustment** to reduce hubness.
- Extend alignment to **diverse language pairs**, including those with different alphabets, without character-level reliance.
- **Enable low-resource language translation** with limited or no parallel data, e.g., English-Esperanto.
- Ensure mapping quality with an **unsupervised model-selection criterion**.

Original Idea

Papers: Mikolov et al. (2013), Xing et al. (2015), Zhang et al. (2017).

- Consider **two sets of independently trained word embeddings**, $X = \{x_1, \ldots, x_n\}$ (source) and $Y = \{y_1, \ldots, y_m\}$ (target).
- Objective: Learn a **linear mapping** W to align translations by minimizing:

$$W^* = \underset{W}{\operatorname{argmin}} ||WX - Y||_F$$

- Translation for a source word s is done by maximizing cosine similarity:

$$t = \underset{t}{\operatorname{argmax}} \cos(Wx_s, y_t)$$

Adversarial Training

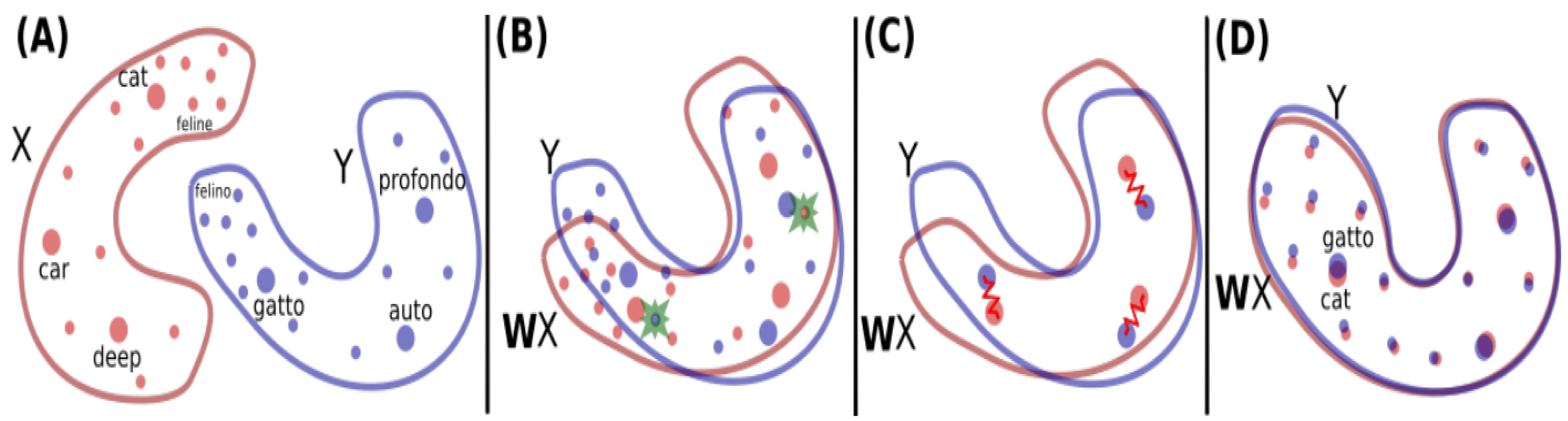
- Adversarial **training aligns** WX **and** Y without cross-lingual supervision.
- **Discriminator** objective: **classify embeddings** as source (transformed WX) or target (Y):

$$L_D(\theta_D|W) = -\frac{1}{n}\sum_{i=1}^{n} \log P_{\theta_D}(source = 1|Wx_i) - \frac{1}{m}\sum_{i=1}^{m} \log P_{\theta_D}(source = 0|y_i)$$

- **Mapping** W objective: **confuse the discriminator**:

$$L_W(W|\theta_D) = -\frac{1}{n}\sum_{i=1}^{n} \log P_{\theta_D}(source = 0|Wx_i) - \frac{1}{m}\sum_{i=1}^{m} \log P_{\theta_D}(source = 1|y_i)$$

Method Illustration



(A) Two distributions of word embeddings are shown: **English words** in red, X, and **Italian words** in blue, Y, which we aim to **align**. Each dot represents a word, sized proportionally to its **frequency** in the training corpus.
(B) Through **adversarial learning**, we learn a **rotation matrix** W that **aligns** the distributions. Green stars represent **randomly selected words** fed to the **discriminator**, which determines if the embeddings come from the same distribution.
(C) The mapping W is **refined via Procrustes**, using **frequent words** from the previous alignment as **anchor points** to minimize an **energy function**, akin to a spring system.
(D) **Translation** uses W and a **CSLS distance metric** that **expands dense areas** (e.g., around "cat") to **reduce "hubness"**, making hubs like "cat" less close to other words compared to panel (A) (Figure 1, p. 3).

Refinement Procedure with Procrustes

- **Initial mapping** W aligns well but **struggles with rare words**.
- **Procrustes refinement** improves accuracy by enforcing **orthogonality**:

$$W^* = \underset{W \in O_d(\mathbb{R})}{\operatorname{argmin}} ||WX - Y||_F = UV^T,$$

where U and V^T are from SVD(YX^T).

- **Frequent words** as **anchors** build a high-quality dictionary.
- **Procrustes** is applied **iteratively** to refine W.

Cross-Domain Similarity Local Scaling (CSLS)

- **CSLS reduces the effect of "hubs"** in dense areas, where some vectors are nearest neighbors for many others.
- **CSLS similarity measure**:

$$CSLS(Wx_s, y_t) = 2\cos(Wx_s, y_t) - r_T(Wx_s) - r_S(y_t)$$

Where:

$$r_T(Wx_s) = \frac{1}{K}\sum_{y_t \in N_T(Wx_s)} \cos(Wx_s, y_t),$$

$$r_S(y_t) = \frac{1}{K}\sum_{x_s \in N_S(y_t)} \cos(x_s, y_t)$$

- **CSLS adjusts similarity based on word density, improving translation accuracy.**

Training and Architectural Choices

- **Word Embeddings:** FastText embeddings with 300 dimensions, trained on Wikipedia; only the top 200k lowercased words.
- **Mapping** W**:** A 300x300 matrix aligning source and target embeddings.
- **Discriminator:** MLP with two 2048-unit layers, Leaky-ReLU activation, 10% dropout, and smoothing $s = 0.2$.
- **Training Procedure:** Discriminator is fed top 50,000 words only; orthogonal updates for stability.
- **Orthogonality Constraint:** Update rule $W \leftarrow (1+\beta)W - \beta(WW^T)W$ with $\beta = 0.01$.
- **Dictionary Generation:** CSLS-selected mutual nearest neighbors boost translation accuracy.
- **Validation:** CSLS-based criterion, correlates with translation accuracy.

Results

- **Procrustes - CSLS (supervised)**: Achieves top P@1 scores, e.g., **81.4 (en-es)**, **82.9 (es-en)**, **72.4 (de-en)**, outperforming other supervised methods (Mikolov et al. (2013), Smith et al. (2017)).
- **Adv - Refine - CSLS (unsupervised)**: Nearly matches Procrustes - CSLS with **81.7 (en-es)**, **83.3 (es-en)**, **74.0 (en-de)**, often surpassing supervised methods.
- **English-Esperanto BLEU Scores**: **NN**: 6.1 (en-eo), 11.9 (eo-en). **CSLS**: 11.1 (en-eo), 14.3 (eo-en), showing clear CSLS improvements.
- **Summary**: Both **Procrustes - CSLS** and **Adv - Refine - CSLS** outperform older methods, with **Adv - Refine - CSLS** highly competitive even without supervision.

	P@1	**P@5**	**P@10**
Methods with cross-lingual supervision			
Mikolov et al. (2013b)†	10.5	18.7	22.8
Dinu et al. (2015)†	45.3	72.4	80.7
Smith et al. (2017)†	54.6	72.7	78.2
Procrustes - NN	42.6	54.7	59.0
Procrustes - CSLS	**66.1**	77.1	80.7
Methods without cross-lingual supervision			
Adv - CSLS	42.5	57.6	63.6
Adv - Refine - CSLS	65.9	**79.7**	**83.1**

	P@1	**P@5**	**P@10**
Methods with cross-lingual supervision			
Mikolov et al. (2013b)†	12.0	22.1	26.7
Dinu et al. (2015)†	48.9	71.3	78.3
Smith et al. (2017)†	42.9	62.2	69.2
Procrustes - NN	53.5	65.5	69.5
Procrustes - CSLS	**69.5**	**79.6**	**83.5**
Methods without cross-lingual supervision			
Adv - CSLS	47.0	62.1	67.8
Adv - Refine - CSLS	69.0	**79.7**	83.1

Table: English to Italian (1), Italian to English (2) word translation retrieval performance (P@1, P@5, P@10) for various methods with and without cross-lingual supervision, evaluated using P@k from 2,000 source queries and 200,000 target sentences. Embeddings from Smith et al. (2017). Results marked by † are theirs. Table 3, p. 8.

References

- Conneau, A., Lample, G., Ranzato, M. A., Denoyer, L., & Jégou, H. (2018). Word translation without parallel data. arXiv. `https://doi.org/10.48550/arXiv.1710.04087`. Code implementation: `https://github.com/facebookresearch/MUSE`
- Dinu, G., Lazaridou, A., & Baroni, M. (2015). Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the International Conference on Learning Representations, Workshop Track*.
- Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. arXiv. `https://doi.org/10.48550/arXiv.1309.4168`.
- Smith, S. L., Turban, D. H. P., Hamblin, S., & Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of the International Conference on Learning Representations*.
- Zhang, M., Liu, Y., Luan, H., & Sun, M. (2017). Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*.
- Xing, C., Wang, D., Liu, C., & Lin, Y. (2015). Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of NAACL*.