




# Word Translation without Parallel Data



Alexis Conneau<sup>z</sup>, Guillaume Lample<sup>x</sup>,  
Marc'Aurelio Ranzato<sup>y</sup>, Ludovic Denoyer<sup>x</sup>, Hervé  
Jégou<sup>y</sup>



# Previous Works

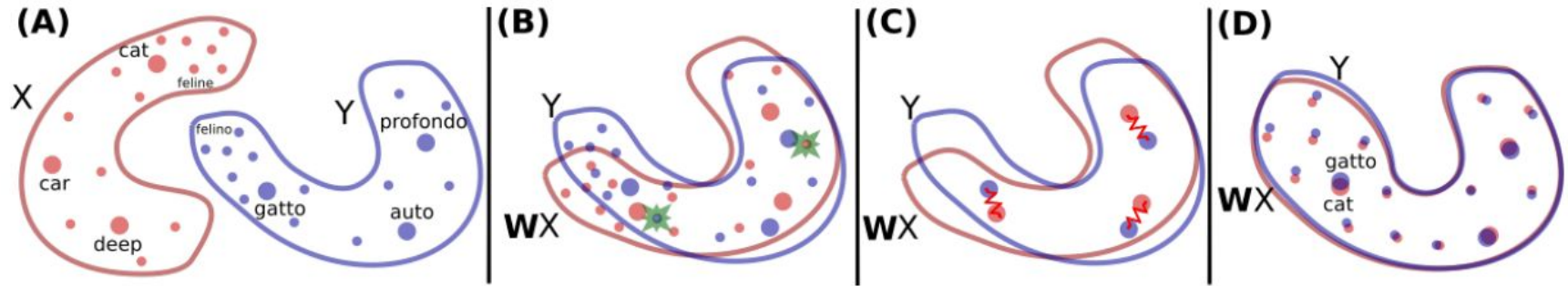
---

- Word Translation with parallel corpora or bilingual dictionaries
- Character level information
- Continuous word embedding spaces exhibit similar structures across languages, even when considering distant language pairs like English and Vietnamese (Mikolov et al.)
  - Used a parallel vocabulary of 5000 words as anchor points
- Zhang et al. used Adversarial Model to obtain cross-lingual word embeddings without any parallel data
  - Performance significantly low than supervised methods

# Quick Overview

---

- Learn word embeddings (FastText) separately on each language using lots of monolingual data.
- Learn a rotation matrix to roughly align the two domains.
- Iterative refinement via orthogonal Procrustes, using the most frequent words.
- Build lexicon using metric that compensates for hubness.



**Figure 1: Toy illustration of the method.** (A) There are two distributions of word embeddings, English words in red denoted by  $X$  and Italian words in blue denoted by  $Y$ , which we want to align/translate. Each dot represents a word in that space. The size of the dot is proportional to the frequency of the words in the training corpus of that language. (B) Using adversarial learning, we learn a rotation matrix  $W$  which roughly aligns the two distributions. The green stars are randomly selected words that are fed to the discriminator to determine whether the two word embeddings come from the same distribution. (C) The mapping  $W$  is further refined via Procrustes. This method uses frequent words aligned by the previous step as anchor points, and minimizes an energy function that corresponds to a spring system between anchor points. The refined mapping is then used to map all words in the dictionary. (D) Finally, we translate by using the mapping  $W$  and a distance metric, dubbed CSLS, that expands the space where there is high density of points (like the area around the word “cat”), so that “hubs” (like the word “cat”) become less close to other word vectors than they would otherwise (compare to the same region in panel (A)).

# Learning Word Embedding

---

Trivial

# Adversarial Training

---

Think of a two player game.

A discriminator is trained to distinguish between the mapped source embeddings and the target embeddings, while the mapping (which can be seen as a generator) is jointly trained to fool the discriminator.

---

Let  $X = \{x_1, x_2, \dots, x_n\}$  and  $Y = \{y_1, y_2, \dots, y_m\}$  be two sets of  $n$  and  $m$  word embeddings.

$X \Rightarrow$  from source and  $Y \Rightarrow$  from target language.

A model is trained to discriminate between elements randomly sampled from  $WX = \{Wx_1, \dots, Wx_n\}$  and  $Y$ . We call this model the discriminator.

$W$  is trained to prevent the discriminator from making accurate predictions.

# Objectives

---

Discriminator Objective:

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

Mapping Objective:

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



# Refinement Procedure

---

Build a synthetic parallel vocabulary using the just learned  $W$  with adversarial training. Only consider the most frequent words and retain only mutual nearest neighbors to ensure a high-quality dictionary.

Apply the Procrustes solution on this generated dictionary iteratively.

# Refinement Contd.

---

Pick most frequent words, translate them via nearest neighbor, solve least square, and iterate.

$x_i$  embedding i-th word in  $E_n$

$y_j$  embedding j-th word in  $I_t$

$$W_t = \arg \min \|W_{t-1}X - Y\|^2, \text{ s.t. } W_t W_t^T = I$$

$W$  orthogonal matrix

# Cross Domain Similarity Local Scaling

---

Nearest Neighbor Relationship => symmetric or asymmetric !!

Expectation:

Nearest neighbor of a source word, in the target language, is likely to have as a nearest neighbor this particular source word.

Reality: (Hubness problem)

Some vectors, are with high probability nearest neighbors of many other points

Others (anti-hubs) are not nearest neighbors of any point.

# CSLS Approach

---

A bi-partite neighborhood graph, in which each word of a given dictionary is connected to its  $K$  nearest neighbors in the other language.

$N_T(Wx_s) \Rightarrow$  the neighborhood, on this bi-partite graph, associated with a mapped source word embedding  $Wx_s$ . All  $K$  elements of  $N_T(Wx_s)$  are words from the target language.

Similarly,  $N_S(y_t) \Rightarrow$  the neighborhood associated with a word  $t$  of the target language.

# CSLS Approach

---

The mean similarity of a source embedding  $x_s$  to its target neighborhood a

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

computed for all source and target word vectors with the efficient nearest neighbors implementation by Johnson et al. (2017).

---

similarity measure CSLS between mapped source words and target words

$$\text{CSLS}(W x_s, y_t) = 2 \cos(W x_s, y_t) - r_T(W x_s) - r_S(y_t).$$

# Orthogonality

---

Keeping the rotation matrix close to a orthogonal one offers several benefits.

- Monolingual quality of the embeddings is preserved.
- An orthogonal matrix preserves the dot product of vectors, as well as their L2 distances, and is therefore an isometry of the Euclidean
- It made the training procedure more stable in our experiments.

$$W \leftarrow (1 + \beta)W - \beta(WW^T)W$$

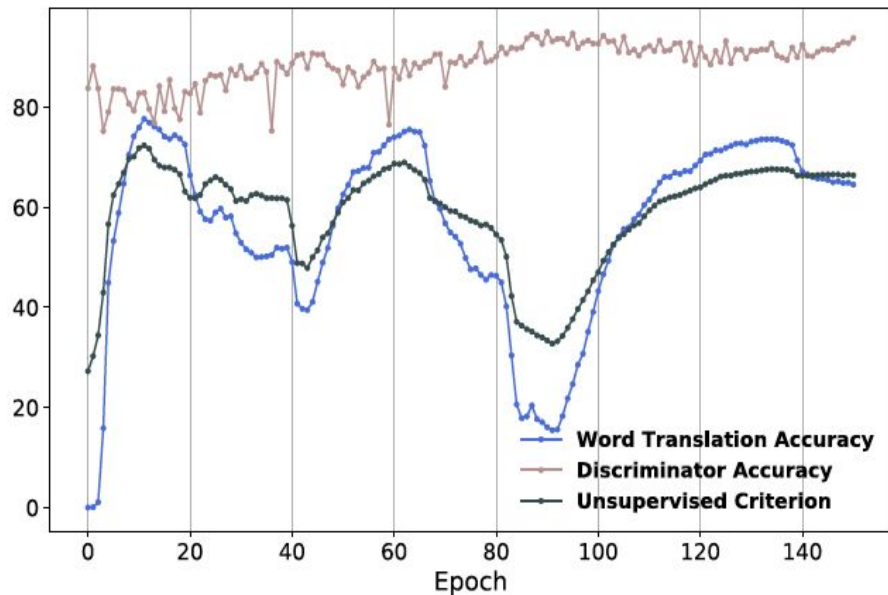
# Unsupervised Criterion

---

- consider 10k most frequent source words, use CSLS to generate translation for each of them and compute average cosine similarity and use this average as validation metric
- this criterion correlates well with the performance of the evaluation task than Wassertein distance



# Unsupervised Criterion



**Figure 2: Unsupervised model selection.** Correlation between our unsupervised validation criterion (black line) and actual word translation accuracy (blue line). In this particular experiment, the selected model is at epoch 10. Observe how our criterion is well correlated with translation accuracy.

# Experiments (Word Translation)

Task considers the problem of retrieving the translation of given source words.  
=>How many times one of the correct translations of a source word is retrieved,  
Precision@k for k = 1; 5; 10.

	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en	en-eo	eo-en
<i>Methods with cross-lingual supervision and fastText embeddings</i>												
Procrustes - NN	77.4	77.3	74.9	76.1	68.4	67.7	47.0	58.2	40.6	30.2	22.1	20.4
Procrustes - ISF	81.1	82.6	81.1	81.3	71.1	71.5	49.5	63.8	35.7	<b>37.5</b>	29.0	27.9
Procrustes - CSLS	81.4	82.9	81.1	<b>82.4</b>	73.5	<b>72.4</b>	<b>51.7</b>	<b>63.7</b>	<b>42.7</b>	36.7	<b>29.3</b>	25.3
<i>Methods without cross-lingual supervision and fastText embeddings</i>												
Adv - NN	69.8	71.3	70.4	61.9	63.1	59.6	29.1	41.5	18.5	22.3	13.5	12.1
Adv - CSLS	75.7	79.7	77.8	71.2	70.1	66.4	37.2	48.1	23.4	28.3	18.6	16.6
Adv - Refine - NN	79.1	78.1	78.1	78.2	71.3	69.6	37.3	54.3	30.9	21.9	20.7	20.6
Adv - Refine - CSLS	<b>81.7</b>	<b>83.3</b>	<b>82.3</b>	82.1	<b>74.0</b>	72.2	44.0	59.1	32.5	31.4	28.2	<b>25.6</b>

**Table 1: Word translation retrieval P@1 for our released vocabularies in various language pairs.** We consider 1,500 source test queries, and 200k target words for each language pair. We use fastText embeddings trained on Wikipedia. NN: nearest neighbors. ISF: inverted softmax. ('en' is English, 'fr' is French, 'de' is German, 'ru' is Russian, 'zh' is classical Chinese and 'eo' is Esperanto)

# Cross Lingual Semantic Word Similarity

---

- A word similarity tasks.
- This task aims at evaluating how well the cosine similarity between two words of different languages correlates with a human-labeled score.
- SemEval 2017 competition data (Camacho-Collados et al. (2017)) which provides large, high quality and well-balanced datasets composed of nominal pairs that are manually scored according to a well-defined similarity scale.
- Report Pearson correlation.

# Cross Lingual Semantic Word Similarity

SemEval 2017	en-es	en-de	en-it
<i>Methods with cross-lingual supervision</i>			
NASARI	0.64	0.60	0.65
our baseline	0.72	0.72	0.71
<i>Methods without cross-lingual supervision</i>			
Adv	0.69	0.70	0.67
Adv - Refine	0.71	0.71	0.71

**Table 4: Cross-lingual wordsim task.** NASARI (Camacho-Collados et al. (2016)) refers to the official SemEval2017 baseline. We report Pearson correlation.

# Sentence translation retrieval

---

- Bag-of words aggregation methods to perform sentence retrieval on the Europarl corpus.
- 2,000 source sentence queries and 200k target sentences for each language pair and report the precision@k for  $k = 1; 5; 10$ , which accounts for the fraction of pairs for which the correct translation of the source words is in the k-th nearest neighbors.
- We use the idf-weighted average to merge word into sentence embeddings.
- The idf weights are obtained using other 300k sentences from Europarl.

# Sentence Translation Retrieval

	en-eo	eo-en
Dictionary - NN	6.1	11.9
Dictionary - CSLS	11.1	14.3

**Table 5: BLEU score on English-Esperanto.**

Although being a naive approach, word-by-word translation is enough to get a rough idea of the input sentence. The quality of the generated dictionary has a significant impact on the BLEU score.



# Sentence Translation Retrieval

Source	mi kelkfoje parolas kun mia najbaro tra la barilo .
Hypothesis	sorry sometimes speaks with my neighbor across the barrier .
Reference	i sometimes talk to my neighbor across the fence .
Source	la viro malanta ili ludas la pianon .
Hypothesis	the man behind they plays the piano .
Reference	the man behind them is playing the piano .
Source	bonvole protektu min kontra tiuj malbonaj viroj .
Hypothesis	gratefully protects hi against those worst men .
Reference	please defend me from such bad men .

**Table 6: Esperanto-English.** Examples of fully unsupervised word-by-word translations. The translations reflect the meaning of the source sentences, and could potentially be improved using a simple language model.

Thank You