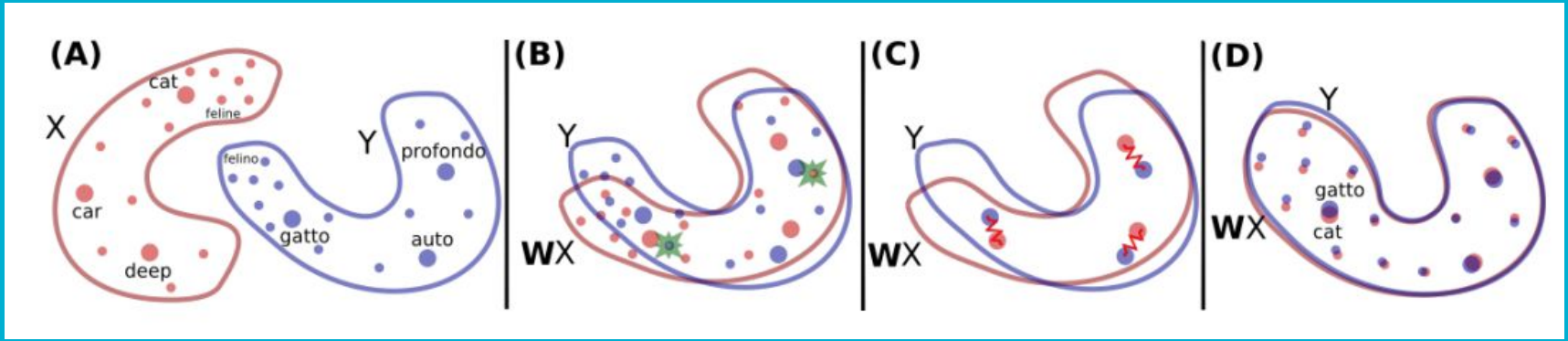
WORD TRANSLATION WITHOUT PARALLEL DATA



Willie Boag

MEDG Reading Group

Overview

- Problem
- Background
- Approach
- Model Selection
- Evaluation

Problem

Given: embedding space for source and embedding space for target

Word-to-word translation

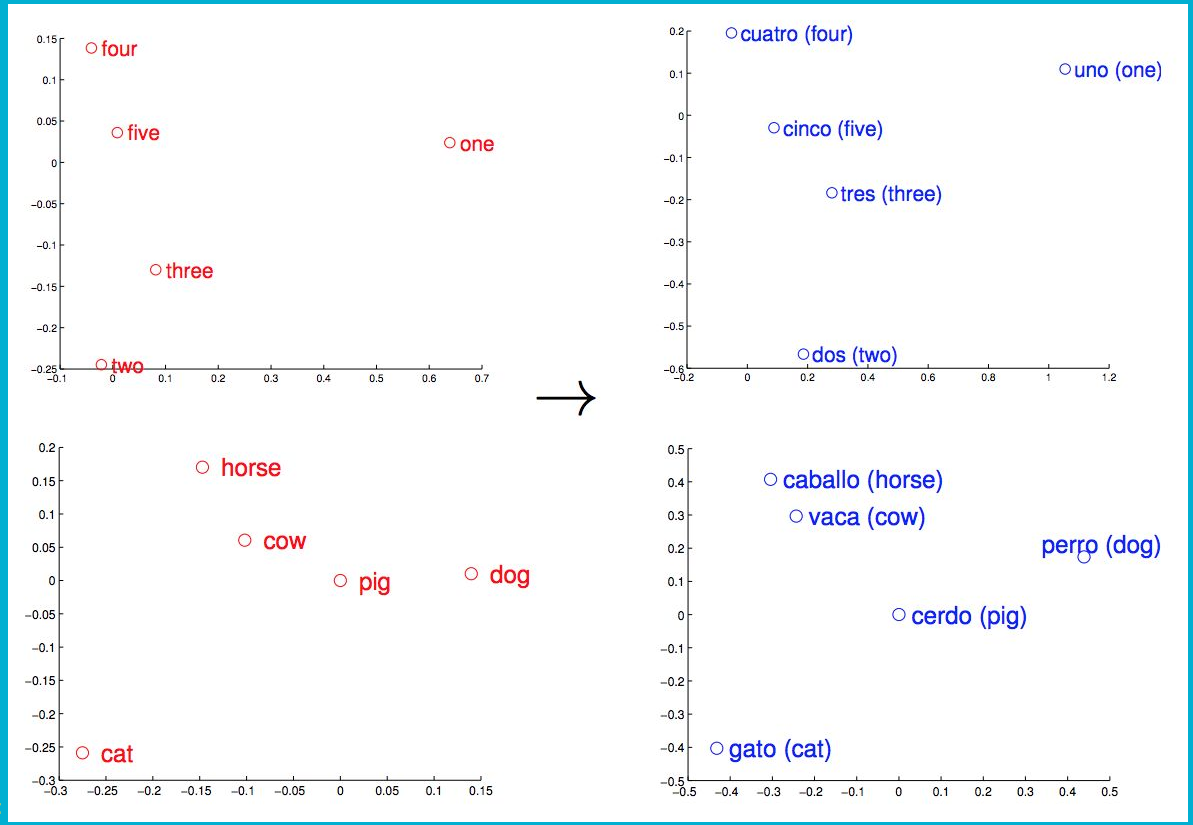
Sometimes, low-resource language pairs (e.g. English-Esperanto)

Background

$$\min_W \sum_{i=1}^{n} \|Wx_i - z_i\|^2$$

Takeaways:

- Supervised
- Linear
- Orthogonal
- In essence “yeah, these embedding spaces are like the same”



https://arxiv.org/pdf/1309.4168v1.pdf

Approach

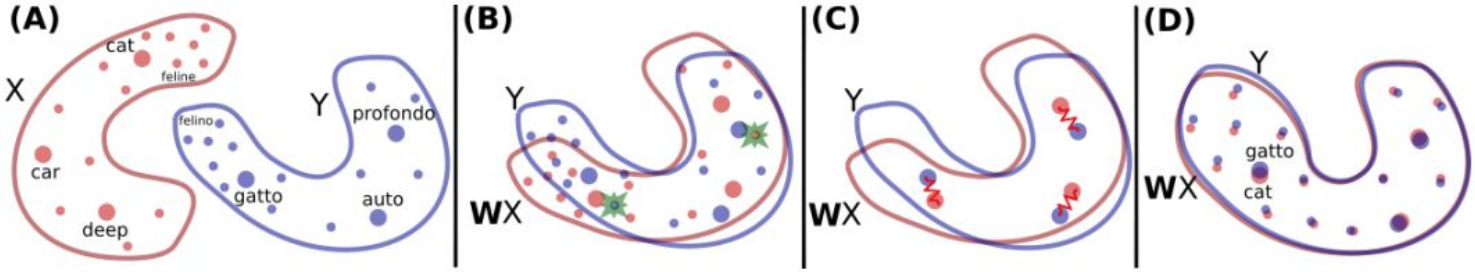
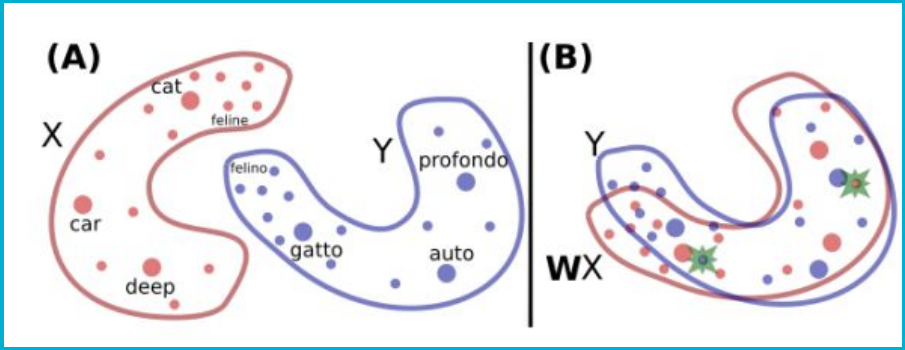


Figure 1: Toy illustration of the method. **(A)** There are two distributions of word embeddings, English words in red denoted by X and Italian words in blue denoted by Y, which we want to align/translate. Each dot represents a word in that space. The size of the dot is proportional to the frequency 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)).

Domain-Adversarial Setting



$P_{\theta_D}(\text{source} = 1|z)$ that a vector z is the mapping of a source embedding

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

Refinement Procedure

1. Generate a synthetic parallel corpus (anchor points)
    a. We consider the most frequent words and retain only mutual nearest neighbors to ensure a high-quality dictionary.

    b. Different metrics one could use for deciding nearest neighbors for synthetic parallel corpus (cosine vs CSCL)

2. Learn mapping from source anchors to target anchors (Procrustes)
3. Apply mapping to all points

Cross-Domain Similarity Local Scaling (CSCL)

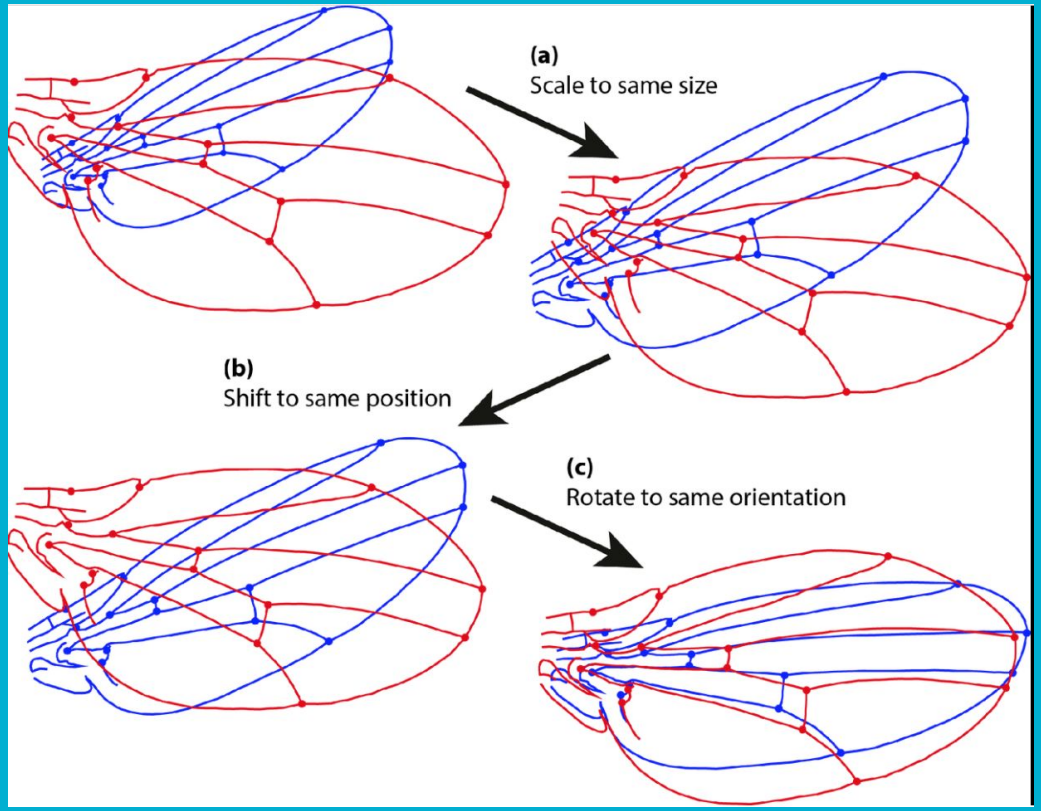
Problem: “hub-ness”

CSLS significantly increases the accuracy for word translation retrieval

$$r_{\mathrm{T}}(Wx_s) = \frac{1}{K} \sum_{y_t \in \mathcal{N}_{\mathrm{T}}(Wx_s)} \cos(Wx_s, y_t)$$

$$\mathrm{CSLS}(Wx_s, y_t) = 2\cos(Wx_s, y_t) - r_{\mathrm{T}}(Wx_s) - r_{\mathrm{S}}(y_t)$$

Procrustes Algorithm



Model Selection

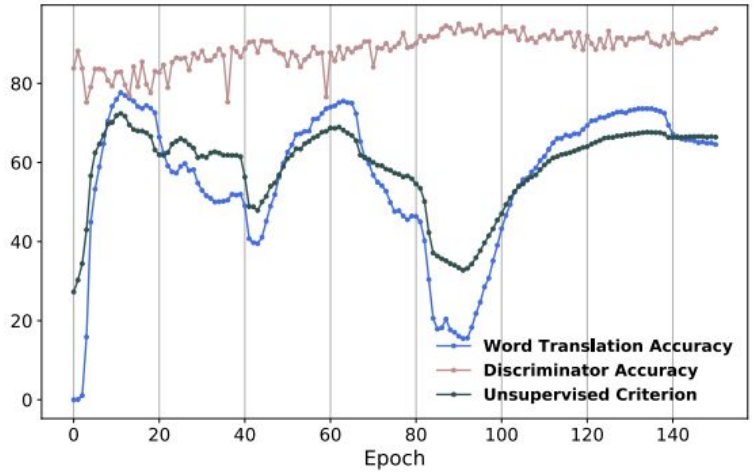


Figure 2: Unsupervised model selection. Correlation between our unsupervised validation criterion (black line) and actual word translation accuracy (blue line). In this particular experiment, the selected model is at epoch 10. Observe how our criterion is well correlated with translation accuracy.

“we consider the 10k most frequent source words, and use CSLS to generate a translation for each of them. We then compute the average cosine similarity between these deemed translations, and use this average as a validation metric”

Evaluation

Word Translation

- Built their own parallel corpora using “an internal translation tool”

Cross-Lingual Semantic Word Similarity

- SemEval 2017: well the cosine similarity between two words of different languages correlates with a human-labeled score

Sentence Translation Retrieval

- But not really (uses BoW sentences)
- Europal corpus with idf-weighted BoW for sentence retrieval

Results

	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en	en-eo	eo-en
Methods with cross-lingual supervision and fastText embeddings												
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	**37.5**	29.0	27.9
Procrustes - CSLS	81.4	82.9	81.1	**82.4**	73.5	**72.4**	**51.7**	**63.7**	**42.7**	36.7	**29.3**	25.3
Methods without cross-lingual supervision and fastText embeddings												
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	**81.7**	**83.3**	**82.3**	82.1	**74.0**	72.2	44.0	59.1	32.5	31.4	28.2	**25.6**

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)

Results

	English to Italian			Italian to English		
	P@1	P@5	P@10	P@1	P@5	P@10
Methods with cross-lingual supervision						
Mikolov et al. (2013b) †	33.8	48.3	53.9	24.9	41.0	47.4
Dinu et al. (2015)†	38.5	56.4	63.9	24.6	45.4	54.1
CCA†	36.1	52.7	58.1	31.0	49.9	57.0
Artetxe et al. (2017)	39.7	54.7	60.5	33.8	52.4	59.1
Smith et al. (2017)†	43.1	60.7	66.4	38.0	58.5	63.6
Procrustes - CSLS	44.9	61.8	66.6	38.5	57.2	63.0
Methods with cross-lingual supervision (Wiki)						
Procrustes - CSLS	63.7	78.6	81.1	56.3	76.2	80.6
Methods without cross-lingual supervision (Wiki)						
Adv - Refine - CSLS	**66.2**	**80.4**	**83.4**	**58.7**	**76.5**	**80.9**

Table 2: English-Italian word translation average precisions (@1, @5, @10) from 1.5k source word queries using 200k target words. Results marked with the symbol † are from Smith et al. (2017). Wiki means the embeddings were trained on Wikipedia using fastText. Note that the method used by Artetxe et al. (2017) does not use the same supervision as other supervised methods, as they only use numbers in their initial parallel dictionary.

Results

	English to Italian			Italian to English		
	P@1	P@5	P@10	P@1	P@5	P@10
Methods with cross-lingual supervision						
Mikolov et al. (2013b) †	10.5	18.7	22.8	12.0	22.1	26.7
Dinu et al. (2015) †	45.3	72.4	80.7	48.9	71.3	78.3
Smith et al. (2017) †	54.6	72.7	78.2	42.9	62.2	69.2
Procrustes - NN	42.6	54.7	59.0	53.5	65.5	69.5
Procrustes - CSLS	**66.1**	77.1	80.7	**69.5**	**79.6**	**83.5**
Methods without cross-lingual supervision						
Adv - CSLS	42.5	57.6	63.6	47.0	62.1	67.8
Adv - Refine - CSLS	65.9	**79.7**	**83.1**	69.0	**79.7**	83.1

Table 3: English-Italian sentence translation retrieval. We report the average P@k from 2,000 source queries using 200,000 target sentences. We use the same embeddings as in Smith et al. (2017). Their results are marked with the symbol †.

Results

SemEval 2017	en-es	en-de	en-it
Methods with cross-lingual supervision			
NASARI	0.64	0.60	0.65
our baseline	0.72	0.72	0.71
Methods without cross-lingual supervision			
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.

System	English			Farsi			German			Italian			Spanish		
	r	ρ	Final	r	ρ	Final	r	ρ	Final	r	ρ	Final	r	ρ	Final
Luminoso_run2	**0.78**	**0.80**	**0.79**	0.51	0.50	0.50	**0.70**	**0.70**	**0.70**	**0.73**	**0.75**	**0.74**	**0.73**	**0.75**	**0.74**
Luminoso_run1	**0.78**	0.79	**0.79**	0.51	0.50	0.50	0.69	0.69	0.69	**0.73**	**0.75**	**0.74**	**0.73**	**0.75**	**0.74**
QLUT_run1*	**0.78**	0.78	0.78	-	-	-	-	-	-	-	-	-	-	-	-
hhu_run1*	0.71	0.70	0.70	0.54	0.59	0.56	-	-	-	-	-	-	-	-	-
HCCL_run1*	0.68	0.70	0.69	0.42	0.45	0.44	0.58	0.61	0.59	0.63	0.67	0.65	0.69	0.72	0.70
NASARI (baseline)	0.68	0.68	0.68	0.41	0.40	0.41	0.51	0.51	0.51	0.60	0.59	0.60	0.60	0.60	0.60

Results

	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.

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.

Critique

- Maybe this task isn’t very hard
  - Assumption is that all embedding spaces, regardless of language, are sort of basically the same (aside from rotation)
  - For word translation evaluation, they generated a large parallel corpus using “an internal translation tool”