Word translation using word embeddings

Emma Covili, Salomé Papereux, Éloïse Zaghrini

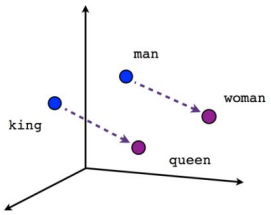
-- DeepL --

Brief statement of the problem

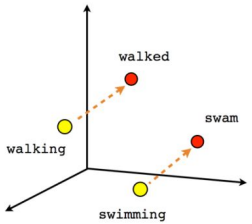
Goal: Build a translator from a source language to a target language

ex: fr→ en, fr→ it ...

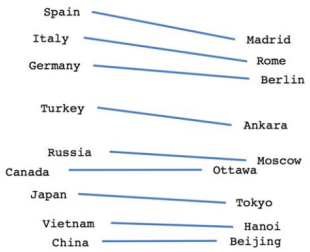
- Idea of the translation :
  - Words used in the same context usually have the same meaning
  - Words can be represented as vectors → Words can be **embedded**



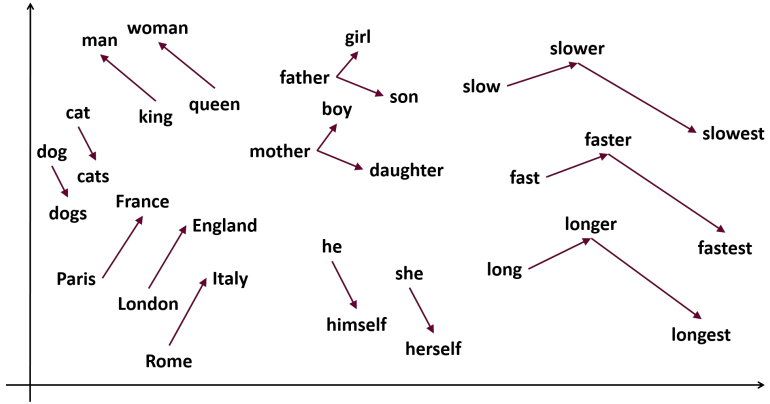
Male-Female



Verb tense

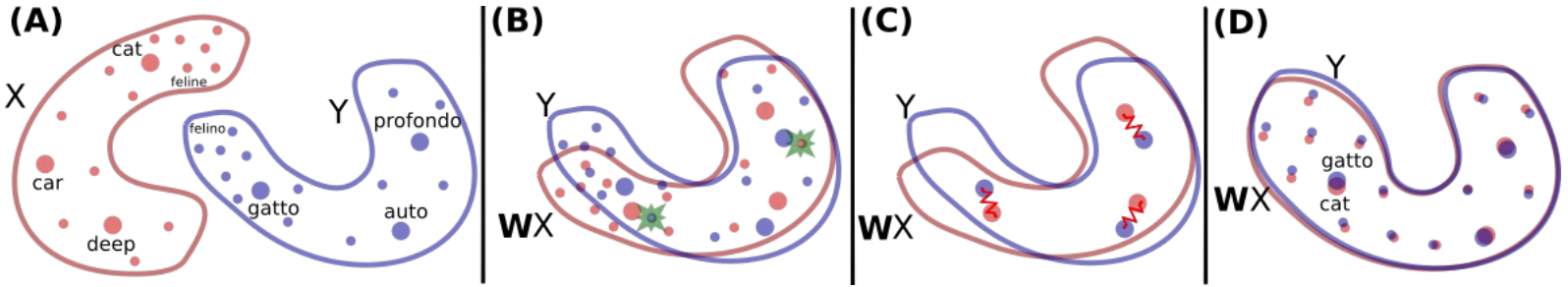


Country-Capital



Methods

X: set of english word embeddings
Y: set of italian word embeddings



We need to compute the matrix W that does the translation between the sets

2 ways to do so:

- Supervised, where we need a bilingual dictionary to build W
- Unsupervised, where W is computed by a GAN

Supervised

Settings: set of pair of words and their associated vector representation $\{x_i, z_i\}_{i=1}^{n}$, where x_i is the vector of the word i in the source language and z_i the vector of its translation.

Linear:

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

Solution of this optimization problem (with regularization) is :

$$W = (XX^T - \lambda Id)X^T Z$$

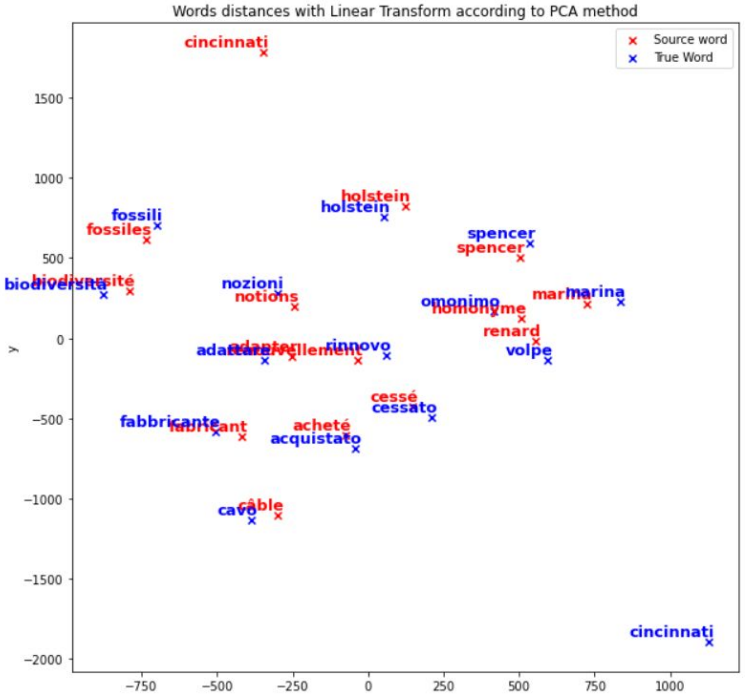
Orthogonal:

$$\begin{cases} \min\limits_{W} \sum\limits_{i=1}^{n} \|Wx_i - z_i\|^2 \\ \min\limits_{\bar{W}} \|W - \bar{W}\| \;\; s.t. \;\; \bar{W}^T \bar{W} = Id \end{cases}$$

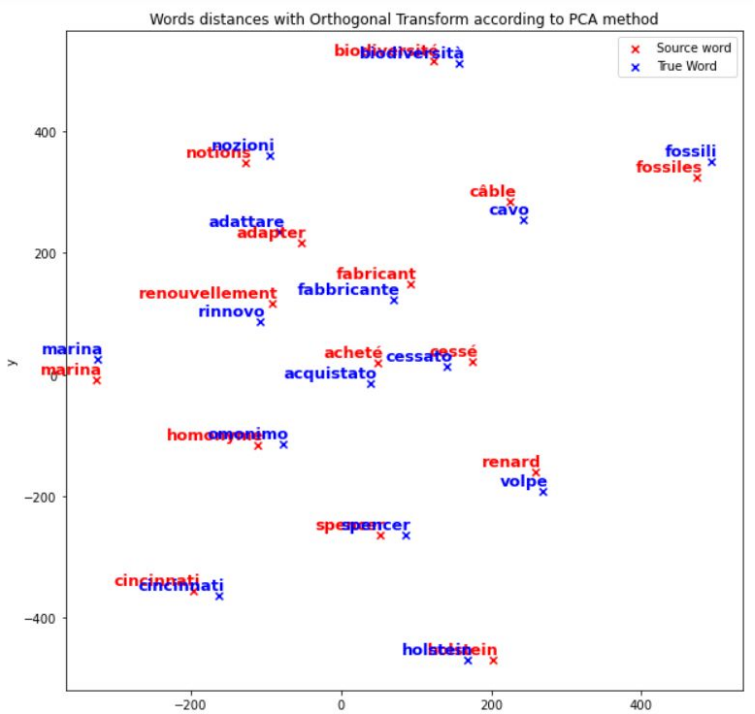
Application of gradient descent and procrustes (through SVD)

Supervised

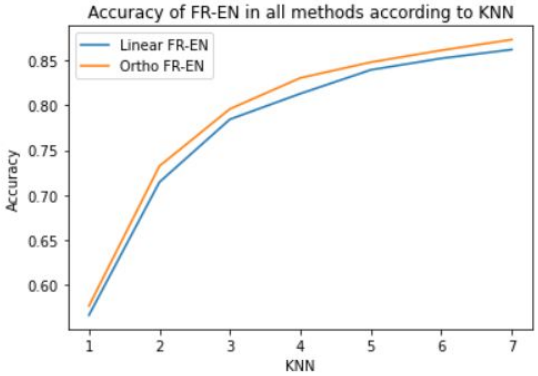
Linear:

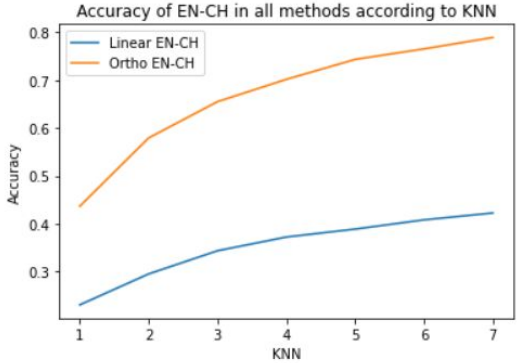


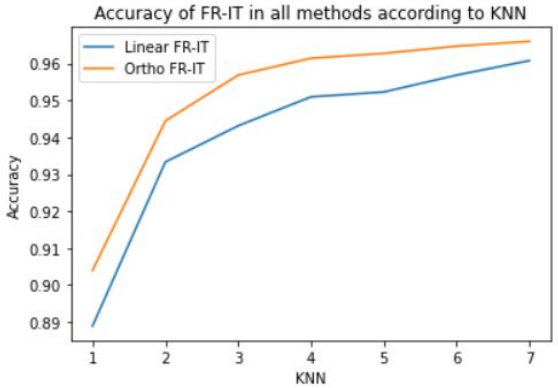
Orthogonal:

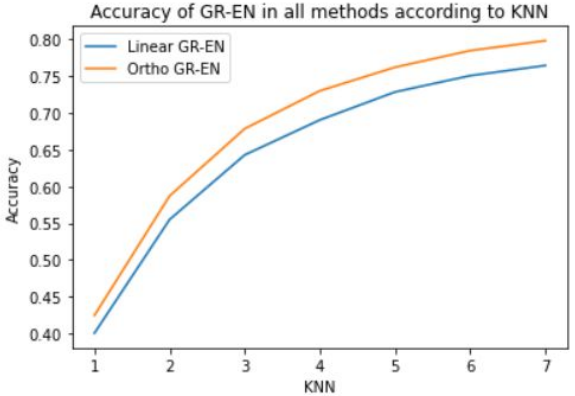


Supervised & KNN









Supervised

	fr-en	en-fr	fr-it	it-fr	en-el	el-en	en-ch
Supervised methods							
Linear - 1-NN	0,57	0,42	0,89	0,78	0,40	0,51	0,23
Linear - 7-NN	0,86	0,87	0,96	0,94	0,76	0,79	0,42
Procrustes - 1-NN	0,57	0,43	0,90	0,79	0,42	0,54	0,43
Procrustes - 7-NN	**0,87**	**0,89**	**0,97**	**0,96**	**0,78**	**0,82**	**0,79**

Note : with the hubness issue resolved, our en-fr translator has an accuracy of 0.57 for 1-NN and 0.72 for 7-NN

Unsupervised

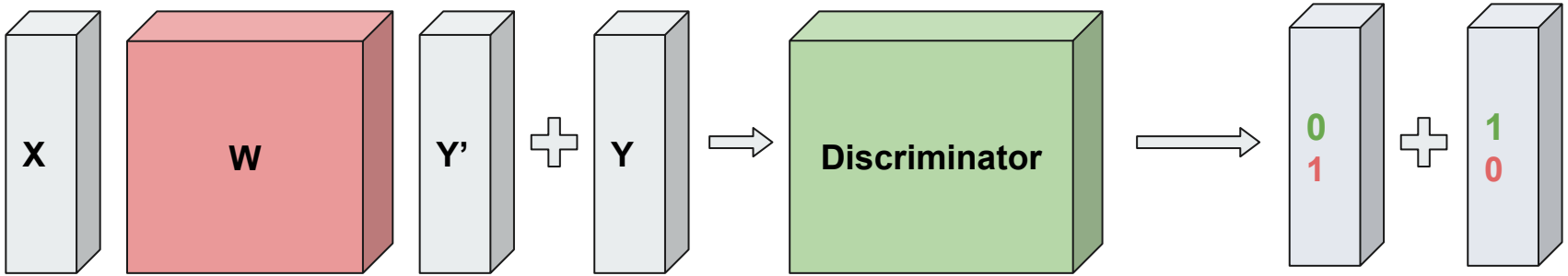
Setting: 2 sets of word embeddings (X, Y) trained independently on monolingual data. The similarities between monolingual embedding spaces can be exploited to learn mappings between the 2 sets with an adversarial model.

Minimization Problem

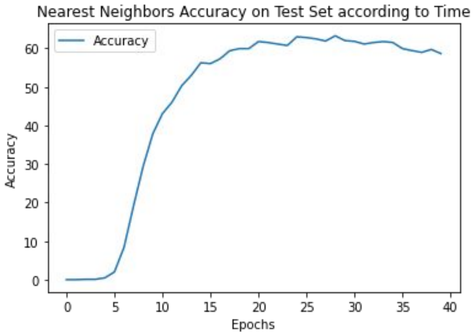
$$W^\star = \underset{W \in M_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_{\mathrm{F}}$$

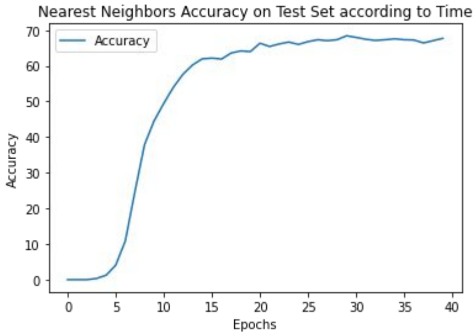
The cosine similarity is used translate words

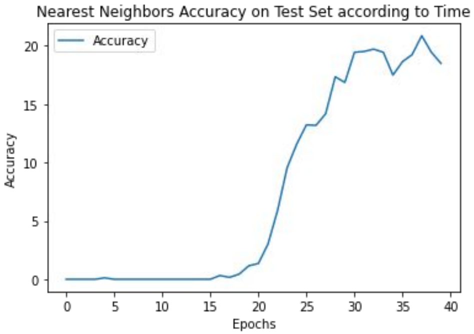
$$t = \operatorname{argmax}_t \cos(W x_s, y_t)$$

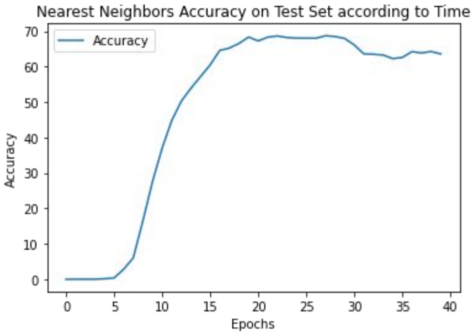


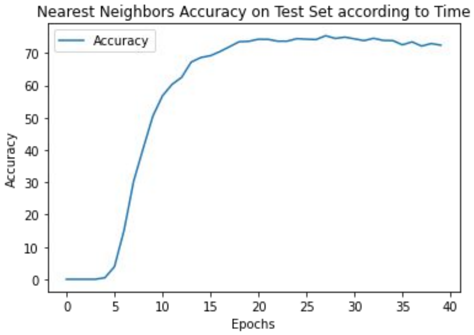
Unsupervised & KNN











1-nn accuracies

	fr-en	en-fr	fr-it	it-fr	en-el	en-ch
linear	0,57	0,42	0,89	0,78	0,40	0,23
procrustes	0,57	0,43	0,90	0,79	0,42	0,43
unsupervised	0,73	0,72	0,86	0,84	0,42	0,45