KØBENHAVNS UNIVERSITET

Comparing Unsupervised Word Translation Methods Step by Step

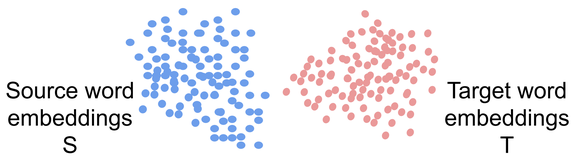
Mareike Hartmann Yova Kementchedjhieva Anders Søgaard
hartmann@di.ku.dk
Department of Computer Science, University of Copenhagen, Denmark

Task: Align monolingual word embeddings such that representations of translations are close to each other

$$W^* = \underset{W}{\operatorname{argmin}} ||WS_d - T_d|| \quad (1)$$

Step 1

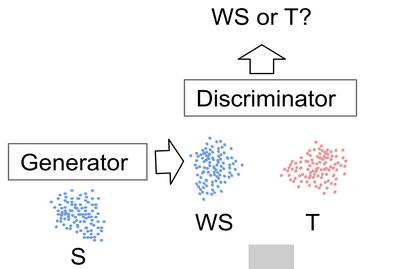
Unsupervised distribution matching



Initial seed dictionary

S_d	T_d
cat	Katze
dog	Pferd
go	gehe
bicycle	Haus
mouse	Maus
...	...

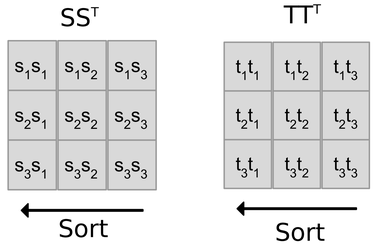
Generative Adversarial Network
(MUSE, Conneau et al. 2018):



Retrieval

S_d T_d

Second order similarities
(VecMap, Artetxe et al. 2018):

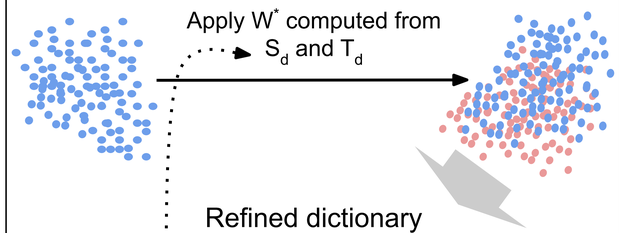


Retrieval

S_d T_d

Step 2

(Pseudo) supervised refinement



Refined dictionary

S_d	T_d
cat	Katze
dog	Hund
go	gehe
bicycle	Haus
mouse	Maus
read	Lesen
dance	Tanz
...	...

Procrustes:
(Schönemann 1966, Conneau et al. 2018)

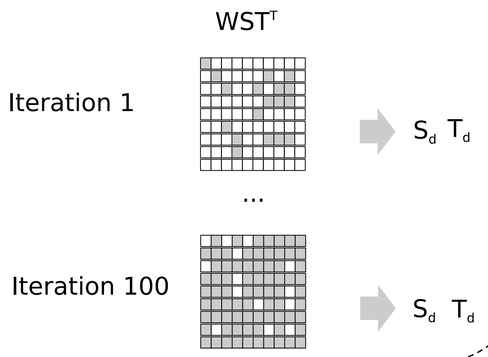
Closed form solution to equation (1)

$$W = UV^T$$

$$S_d T_d^T = U\Sigma V^T$$

Procrustes with Stochastic Dictionary Induction (SDI):
(Artetxe et al. 2018)

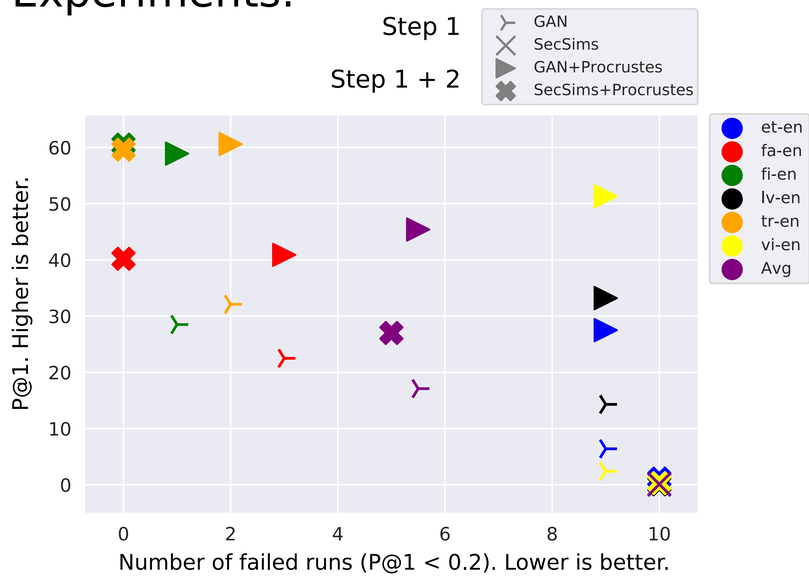
Randomly drop candidates for the refined dictionary



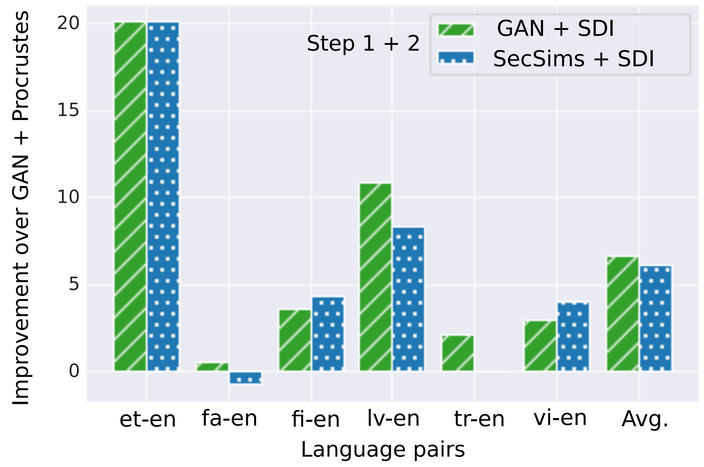
S_d T_d

S_d T_d

Experiments:



Select best of 10 GAN initializations using and unsupervised selection criterion (**csls**, Conneau et al. 2018) and combine with SDI



Take away:

GAN-based initialization is unstable but has the highest potential.