Adversarial Training for Unsupervised Bilingual Lexicon Induction

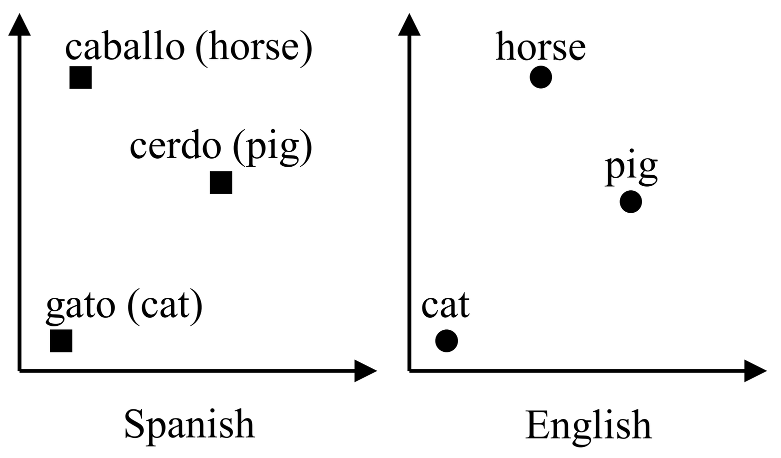
Meng Zhang, Yang Liu, Huanbo Luan, Maosong Sun

Tsinghua University, Beijing, China

Overview

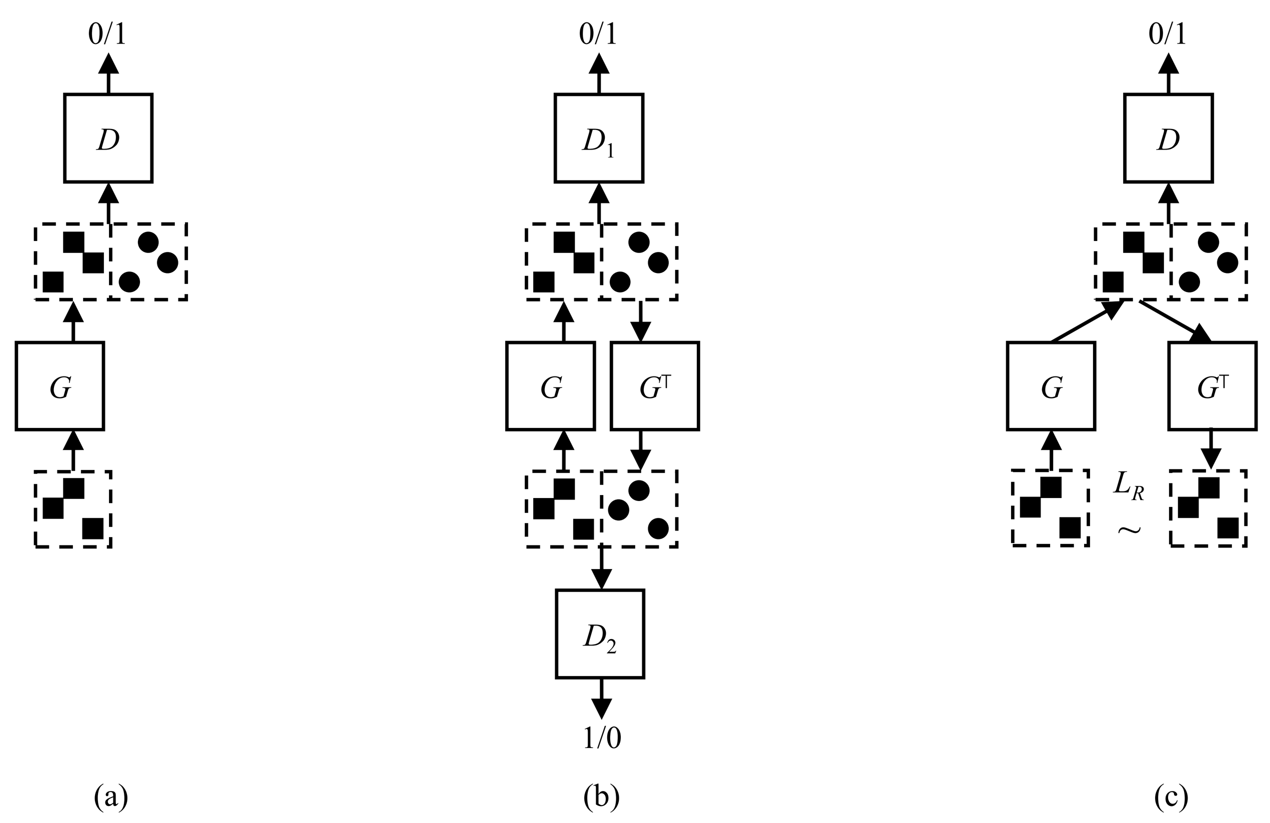
- Task input: Separate monolingual embeddings trained on non-parallel data
- Task output: A bilingual lexicon
- Challenge: Zero supervision: Can we link separate monolingual embeddings without any cross-lingual signal?
- Solution: Formulate as an adversarial game
- Outcome: Successful learning with proper model design and training techniques

Background



Although monolingual word embeddings are trained separately on non-parallel data, they appear approximately isomorphic. Therefore a linear transformation can be used to align the two embedding spaces. But previous works typically require seed word translation pairs to supervise its learning.

Models



(a) Model 1 (unidirectional transformation): The generator G is a linear transformation that tries to transform source word embeddings (squares) to make them seem like target ones (dots), while the discriminator D tries to classify whether the input embeddings are generated by G or real samples from the target embedding distribution.
(b) Model 2 (bidirectional transformation): If G transforms the source word embedding space into the target language space, its transpose $G^\top$ should transform the target language space back to the source.
(c) Model 3 (adversarial autoencoder): After the generator G transforms a source word embedding x into a target language representation Gx, we should be able to reconstruct the source word embedding x by mapping back with $G^\top$.

- Model 2 and 3 can be seen as relaxations of an orthogonal constraint on G.

Training Techniques

Regularizing the discriminator

- All forms of regularization help training.
- Multiplicative Gaussian injected into the input is the most effective. On top of that, hidden layer noise helps slightly.

Model selection

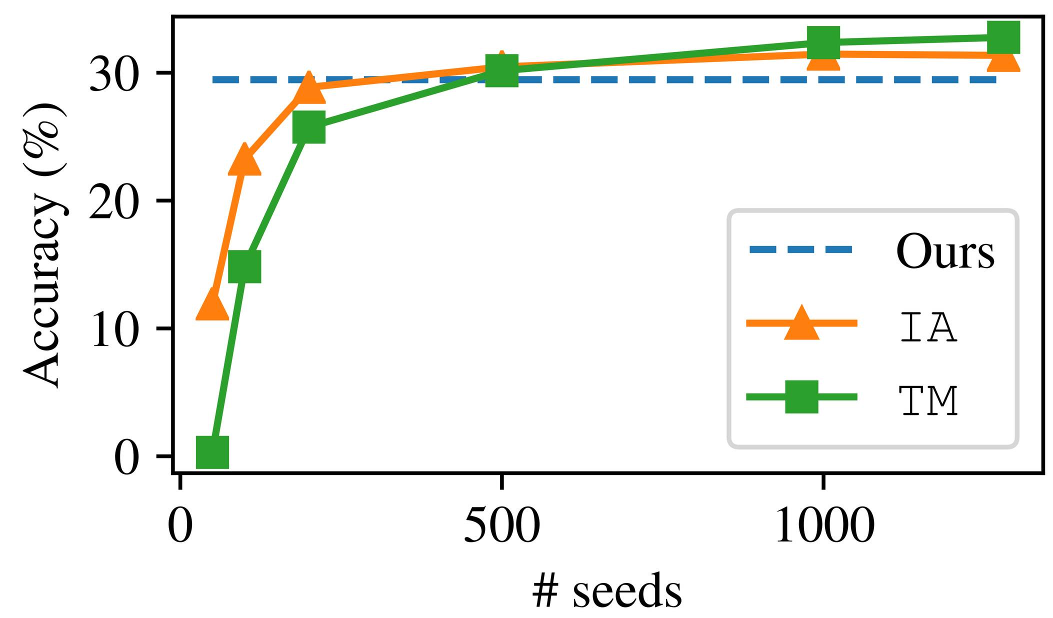
- Sharp drops of the generator loss correspond to good models.
- Reconstruction loss L_R and the value of $\left\|G^\top G - I\right\|_F$ drop synchronously → Good models are indeed close to orthogonality.

Experiments

Chinese-English

method	# seeds	accuracy (%)
MonoGiza w/o emb.	0	0.05
MonoGiza w/ emb.	0	0.09
TM	50	0.29
	100	21.79
IA	50	18.71
	100	32.29
Model 1	0	39.25
Model 1 + ortho.	0	28.62
Model 2	0	40.28
Model 3	0	43.31

Comparison with seed-based methods



Spanish-English, Italian-English, Japanese-Chinese, Turkish-English

method	# seeds	es-en	it-en	ja-zh	tr-en
MonoGiza w/o embeddings	0	0.35	0.30	0.04	0.00
MonoGiza w/ embeddings	0	1.19	0.27	0.23	0.09
TM	50	1.24	0.76	0.35	0.09
	100	48.61	37.95	26.67	11.15
IA	50	39.89	27.03	19.04	7.58
	100	60.44	46.52	36.35	17.11
Ours	0	71.97	58.60	43.02	17.18

Large-scale settings

method	# seeds	Wikipedia	Gigaword
TM	50	0.00	0.01
	100	4.79	2.07
IA	50	3.25	1.68
	100	7.08	4.18
Ours	0	7.92	2.53

Conclusion

- Feasible to connect the word embeddings of different languages without any cross-lingual signal
- Comparable performance with methods that require seeds to train
- Code available at `http://thunlp.org/~zm/UBiLexAT/`