Distributional Inclusion Vector Embedding for Unsupervised Hypernym Detection

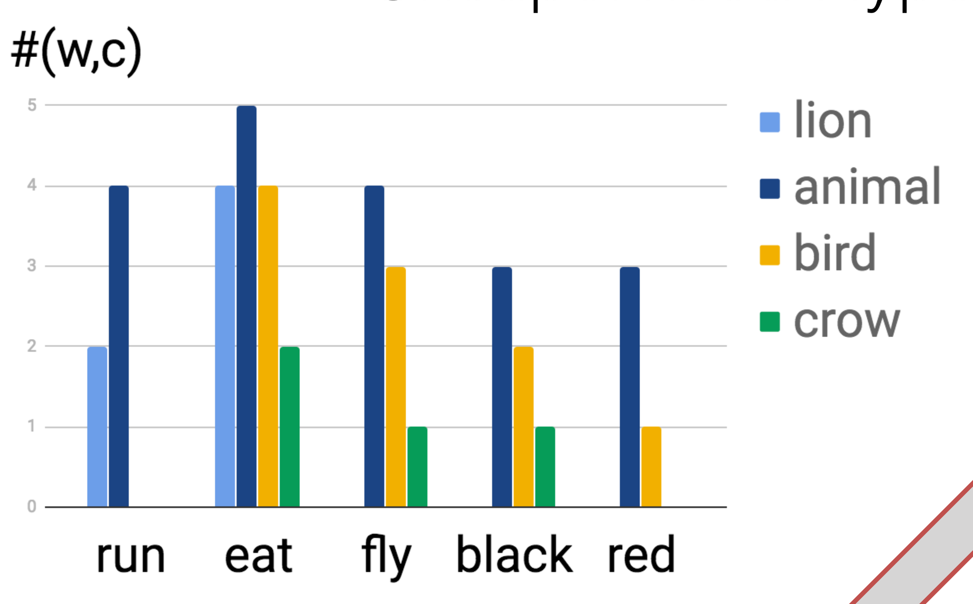
UMASS AMHERST

Haw-Shiuan Chang, ZiYun Wang, Luke Vilnis, Andrew McCallum

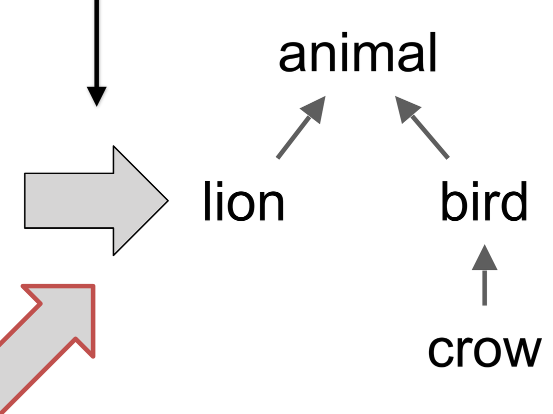
IESL

Introduction

Unsupervised Hypernym Detection [1]



Sparse bag-of-words (SBOW)

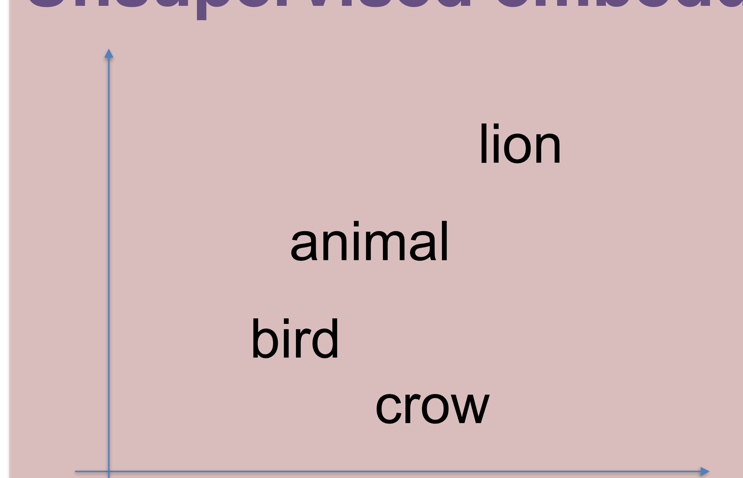


Taxonomy

In Out In Out

Related Work

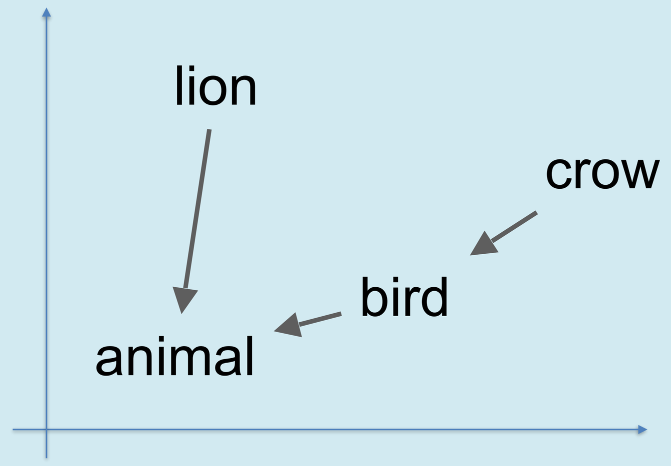
Unsupervised embedding



Word2Vec

Most of unsupervised embeddings do not preserve frequency signal

Supervised embedding

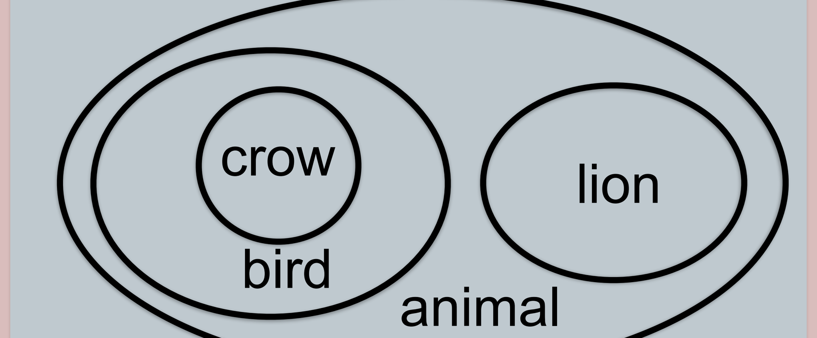


Order embedding [2]

Our method trains order embedding without supervision

Topics	w with top P(w|topic)
run	… lion, … animal…
fly	… bird, … crow…
city	… crow, … bird…

Topic modeling / NMF



Gaussian embedding [3]

Method

Distributional Inclusion Vector Embedding (DIVE)

As efficient as skip-gram

Non-negative

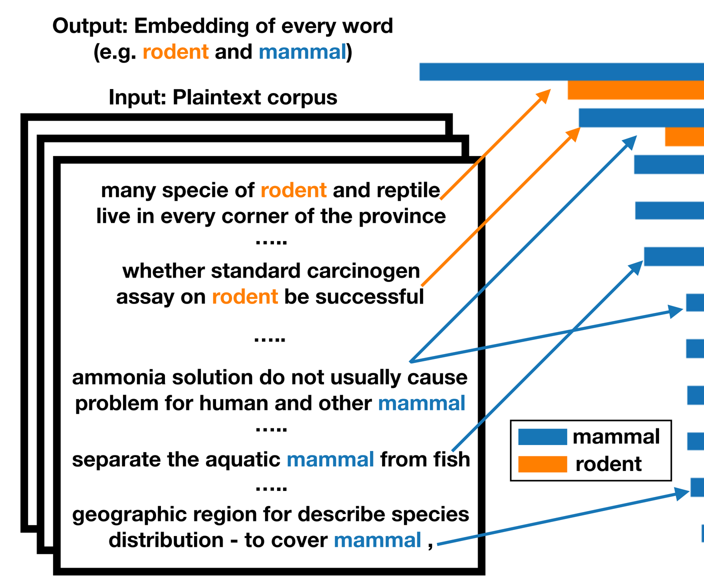
$$l = \sum_{w \in V} \sum_{c} \#(w,c) \log \sigma(\mathbf{w}^T \mathbf{c}) + k \sum_{w \in V} \sum_{c} \#(w,c) \mathop{\mathbb{E}}_{c_N \sim P_D} [\log \sigma(-\mathbf{w}^T \mathbf{c_N})],$$

PMI>k

rodents include rats

* Z/#(w)

mean(#(w))



id	Top 1-5 words	Top 51-55 words
1	find, specie, species, animal, bird	hunt, terrestrial, lion, planet, shark
2	system, blood, vessel, artery, intestine	function, red, urinary, urine, tumor
3	head, leg, long, foot, hand	shoe, pack, food, short, right
4	may, cell, protein, gene, receptor	neuron, eukaryotic, immune, kinase, generally
5	sea, lake, river, area, water	terrain, southern, mediterranean, highland, shallow
6	cause, disease, effect, infection, increase	stress, problem, natural, earth, hazard
7	female, age, woman, male, household	spread, friend, son, city, infant
8	food, fruit, vegetable, meat, potato	fresh, flour, butter, leave, beverage
9	element, gas, atom, rock, carbon	light, dense, radioactive, composition, deposit
10	number, million, total, population, estimate	increase, less, capita, reach, male
11	industry, export, industrial, economy, company	centre, chemical, construction, fish, small

Inclusion Property in Embedding Space

$$x \preceq y \iff \mathbf{x}[i] \leq \mathbf{y}[i] \ , \forall i \in \{1, ..., d_0\}$$

We make **x**, **y**, **c** non-negative in DIVE

For all **c**, $\mathbf{x}^T\mathbf{c} \leq \mathbf{y}^T\mathbf{c}$

$\mathbf{x}^T\mathbf{c} \approx M'[x,c]$ and $\mathbf{y}^T\mathbf{c} \approx M'[y,c]$ (approximately speaking [4])

For all c, $M'[x,c] \leq M'[y,c]$

$M'[x,c] = \log\left(\frac{\#(x,c)|V|}{\#(c)k}\right)$ and $M'[y,c] = \log\left(\frac{\#(y,c)|V|}{\#(c)k}\right)$

Not depend on y

$$x \preceq y \iff \forall c,\ \#(x,c) \leq \#(y,c),$$

Distributional inclusion hypothesis

Experiments

- Trained on the first ~50 million words of wikipedia
- Test on 10 datasets using AP@all (AUC) and HyperLex using ρ

Dataset	BLESS	EVALution	LenciBenotto	Weeds	Medical	LEDS
Random	5.3	26.6	41.2	51.4	8.5	50.5
Word2Vec + C	9.2	25.4	40.8	51.6	11.2	71.8
GE + C	10.5	26.7	43.3	52.0	14.9	69.7
GE + KL	7.6	29.6	45.1	51.3	15.7	64.6 (80^3)
DIVE + C·ΔS	**16.3**	**33.0**	**50.4**	**65.5**	**19.2**	**83.5**

Dataset	TM14	Kotlerman 2010	HypeNet	WordNet	Avg (10 datasets)	HyperLex
Random	52.0	30.8	24.5	55.2	23.2	0
Word2Vec + C	52.1	**39.5**	20.7	**63.0**	25.3	16.3
GE + C	53.9	36.0	21.6	58.2	26.1	16.4
GE + KL	52.0	39.4	23.7	54.4	25.9	9.6 (20.6^3)
DIVE + C·ΔS	**57.2**	36.6	**32.0**	60.9	**32.7**	**32.8**

Distributional Inclusion Generality Similarity * Generality

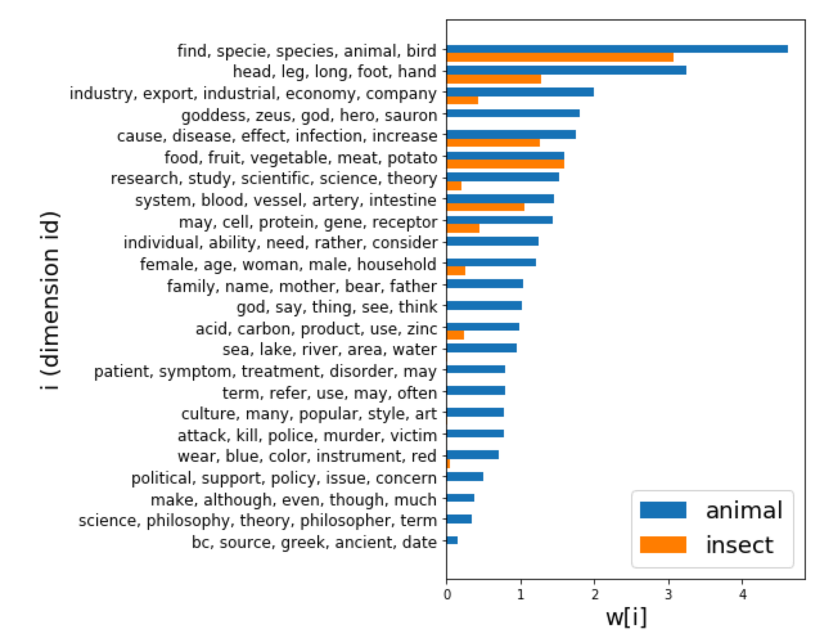
AP@all (%)		Micro Average (10 datasets) CDE	AL_1	ΔS	W·ΔS	C·ΔS
SBOW	Freq	31.1	**28.2**	31.5	31.6	31.2
	PPMI	30.1	23.0	31.1	32.9	**33.5**
	PPMI w/ IS	**31.8**	24.1	31.5	32.1	30.3
	All wiki	29.0	23.1	29.2	30.2	31.1
DIVE	Full	27.6	25.3	**32.1**	**34.1**	32.7
	w/o PMI	28.5	26.7	31.5	33.4	30.1
	w/o IS	22.3	20.7	19.1	19.6	19.9
Kmean (Freq NMF)		29.1	24.7	31.5	31.8	31.5

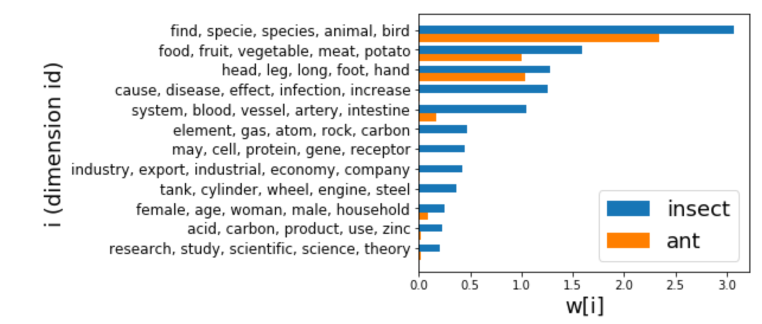
w/o PMI>k

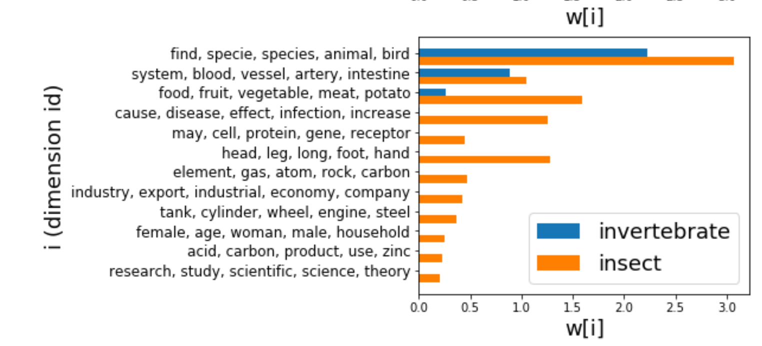
w/o *Z/#(w) (non-negative word2vec)

Average # dimensions

SBOW Freq	5799
SBOW PPMI	3808
DIVE	**20**

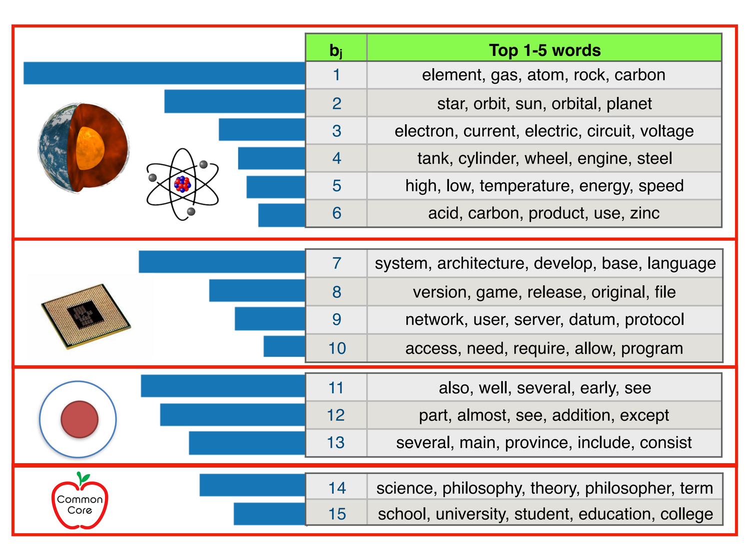




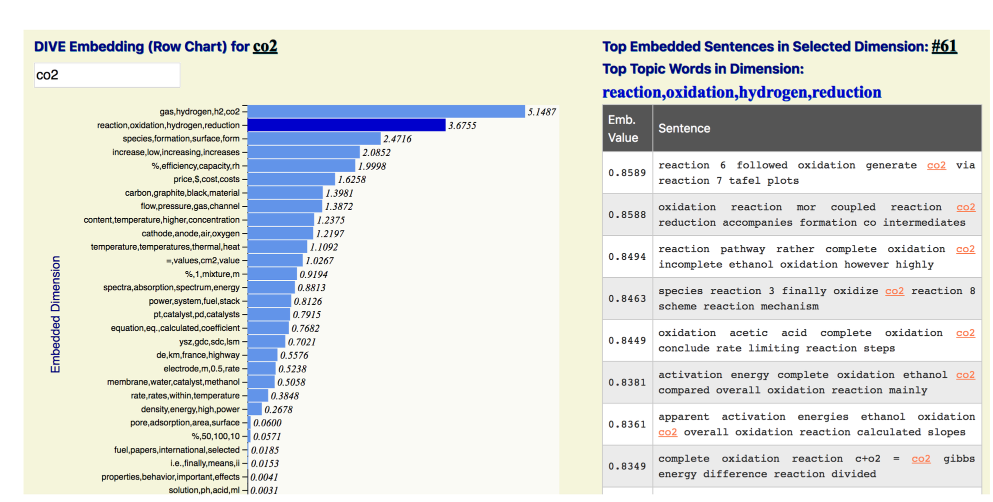


Other Applications

- DIVE is a compression of sparse bag of words (SBOW)
- Compression makes processing more efficient and visualizable



Word sense induction for the word core (TextGraph 2018 [5])



Word co-occurrence statistics visualization for the word CO_2 in material science papers

Conclusion

- We propose an interpretable and efficient word embedding, DIVE
- DIVE compresses SBOW while preserving the inclusion property
- DIVE achieves new state-of-the-art unsupervised performance on most of hypernym detection datasets

References

[1] V. Shwartz, E. Santus, and D. Schlechtweg. 2017. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In EACL.

[2] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. 2016. Order-embeddings of images and language. In ICLR.

[3] L. Vilnis and A. McCallum. 2015. Word representations via gaussian embedding. In ICLR.

[4] O. Levy and Y. Goldberg. 2014. Neural word embedding as implicit matrix factorization. In NIPS.

[5] H.-S. Chang, A. Agrawal, A. Ganesh, A. Desai, V. Mathur, A. Hough, A. McCallum. 2018. Efficient Graph-based Word Sense Induction by Distributional Inclusion Vector Embeddings, In TextGraphs