Language Technology Lab

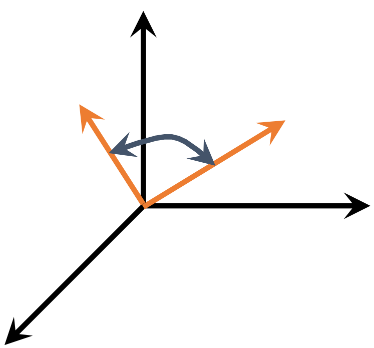
UNIVERSITY OF CAMBRIDGE

Inducing Embeddings for Rare and Unseen Words by Leveraging Lexical Resources

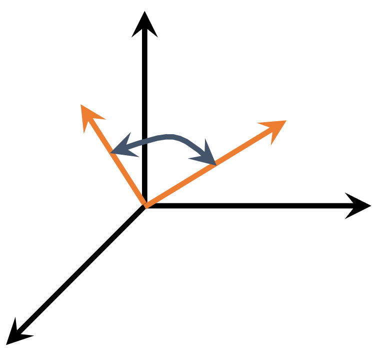
Mohammad Taher Pilehvar

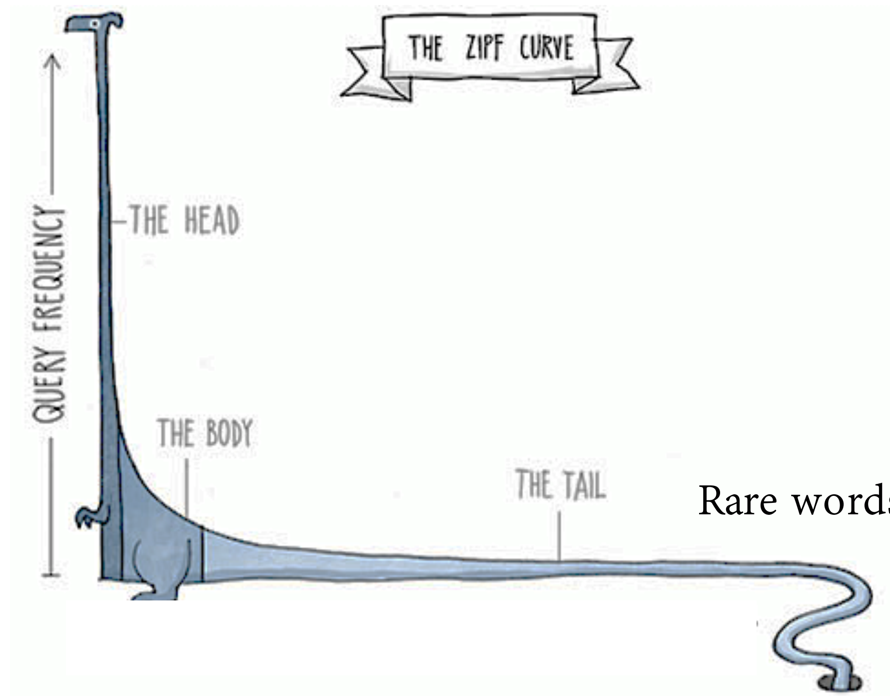
Nigel Collier

Representation of rare words



Representation of rare words





Unseen words

Illustration by Eva-Lotta
www.smashingmagazine.com

Representation of rare words: past work

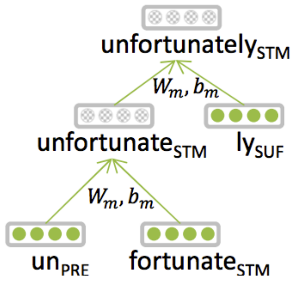
Recent work has mainly focused on

morphologically complex rare words

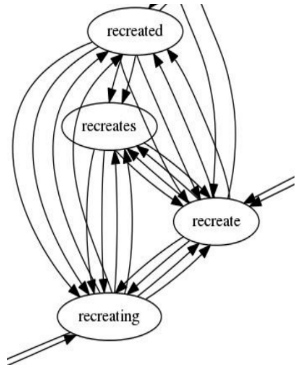
intercommunicate disturbances postmodernism

unconsciousness corroding

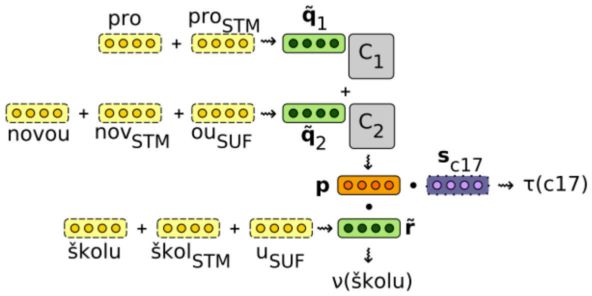
Luong et al. (2013)



Soricut and Och (2015)



Botha and Blunsom (2014)



Representation of rare words

What about other (single morpheme) rare words?

Infrequent or domain-specific words

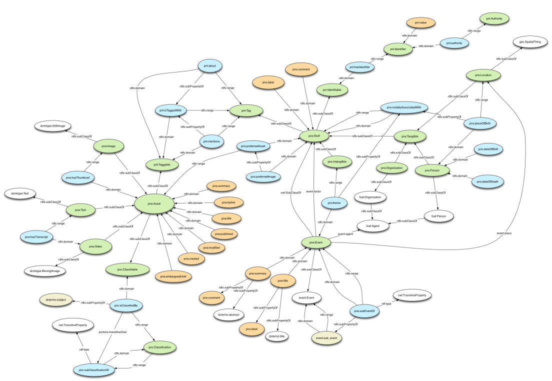
piquance
bronchomegaly
degust
abouphalia
hyperthreoninuria
enthuse
milphosis
schmetterlingswirbel
lieutenant

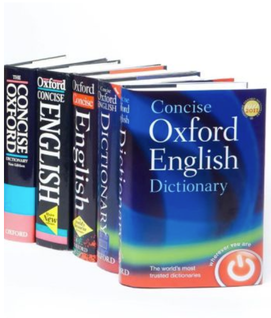
Lexical resources to the rescue

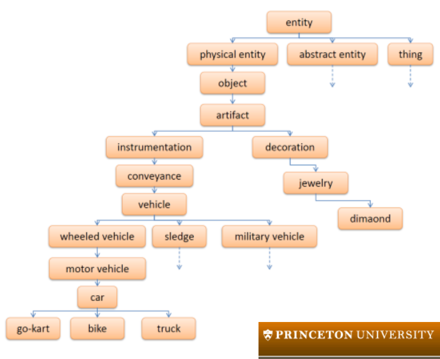
Abundance of domain-specific lexical resources: ontologies, dictionaries, databases, etc.

BabelNet

piquance
bronchomegaly
degust
abouphalia
hyperthreoninuria
enthuse
milphosis
schmetterlingswirbel
lieutenant





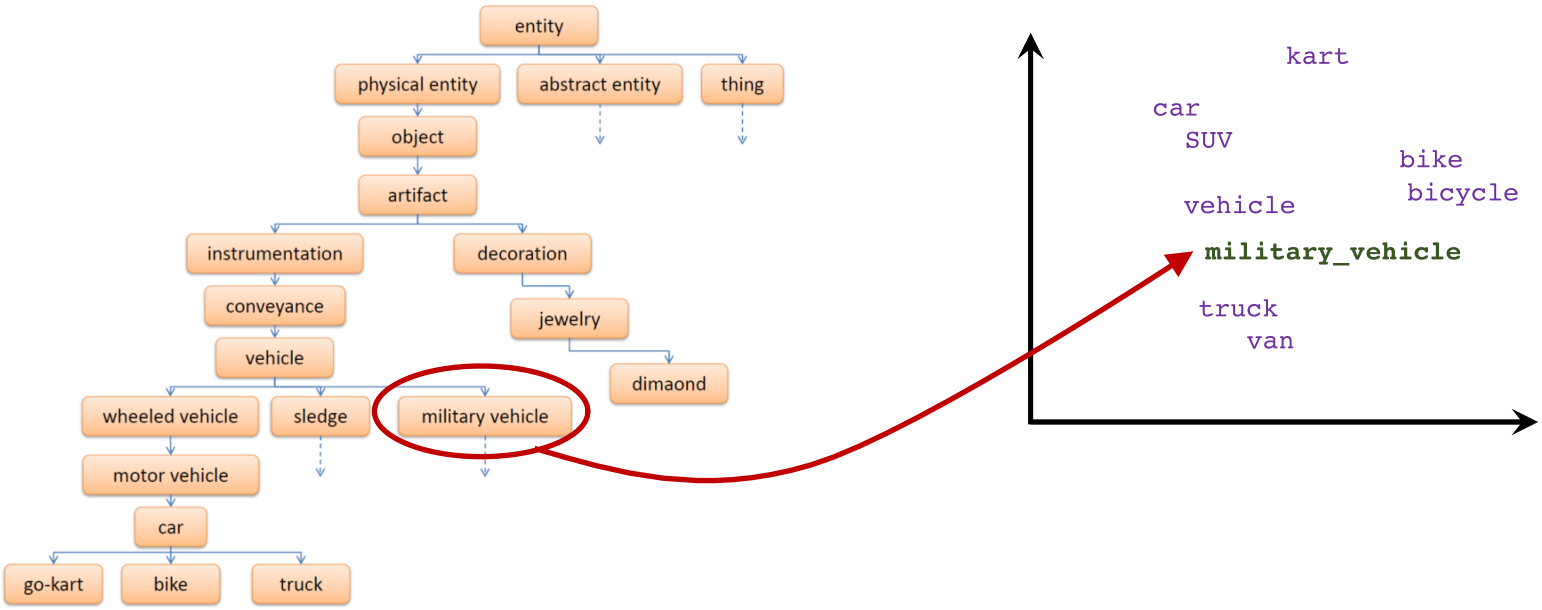


MeSH

PRINCETON UNIVERSITY
WordNet
A lexical database for English

Inducing embedding for unseen words

Leverage knowledge encoded in external lexical resources for inducing embeddings



Inducing embedding for unseen words

The proposed procedure:

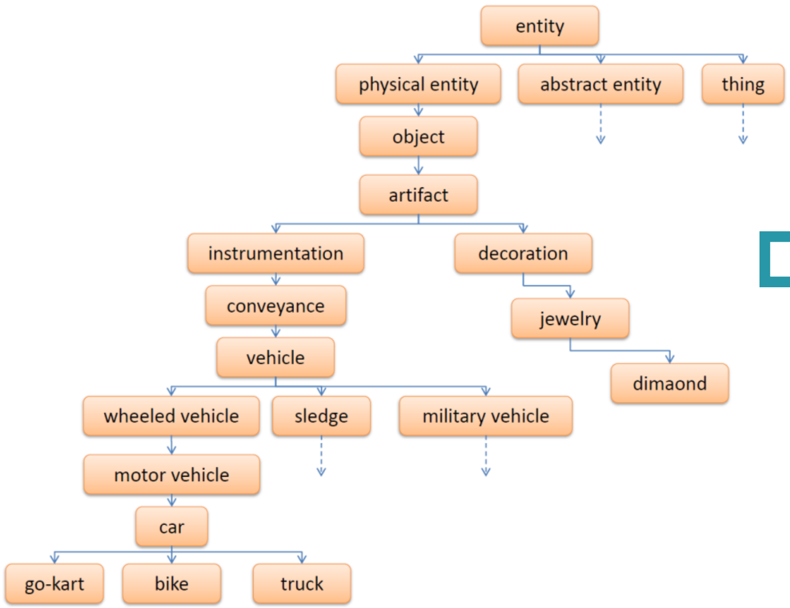
1. View a lexical resource as a **semantic network**

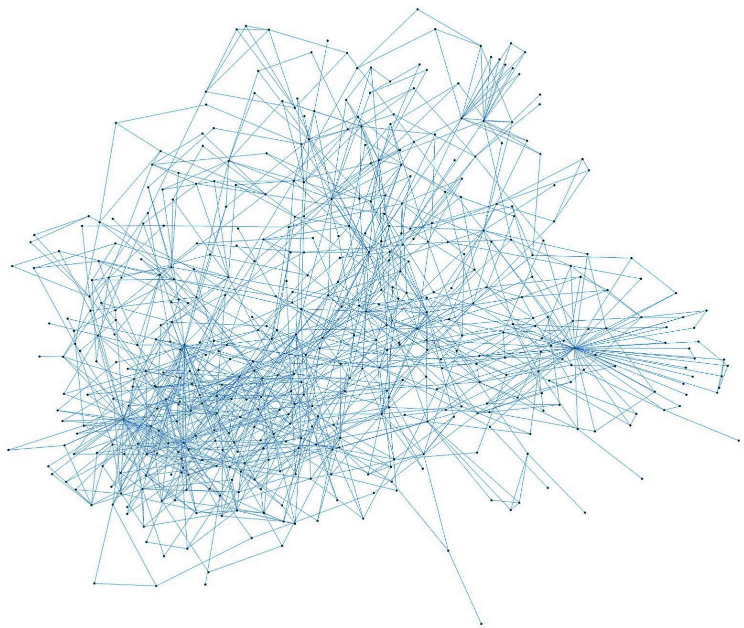
2. Extract for an unseen word its set of **semantic landmarks**

3. **Induce** the embedding for the unseen word using its landmarks

Inducing embedding for unseen words

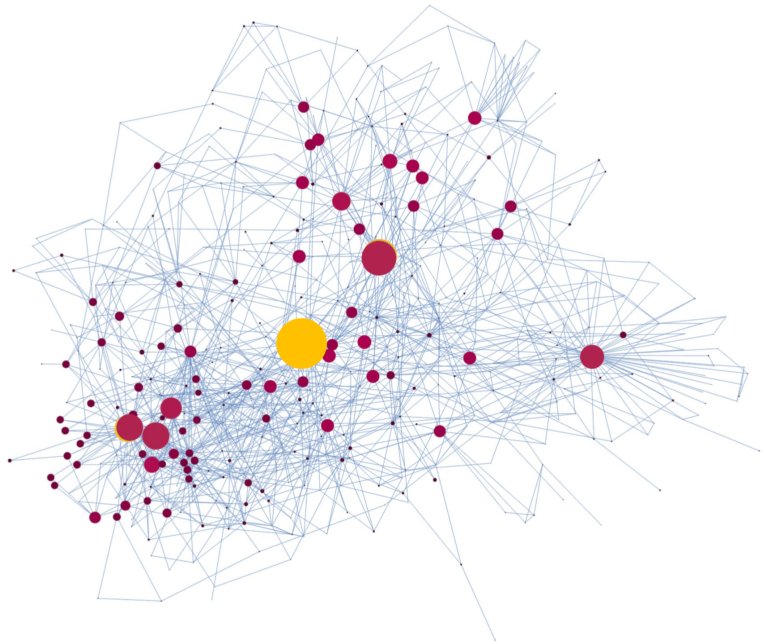
1. View lexical resource as a semantic network





Inducing embedding for unseen words

2. Extract semantic landmarks



Using Personalized PageRank

military_vehicle
vehicle
military_machine
caisson
tank
humvee
troop_carrier
pickup
warplane
Lorry
Warship
picket
personnel

Inducing embedding for unseen words

3. Induce embeddings

semantic landmarks

$$\hat{\mathbf{d}}(w_r) = \theta\,\mathbf{d}(w_r^0) + \frac{1}{\mathcal{L}_r}\sum_{i=1}^{|\mathcal{L}_r|} \mathrm{e}^{-i}\mathbf{d}(l_{i,r})$$

induced embedding for w_r

embedding for the i-th landmark

Inducing embedding for unseen words

3. Induce embeddings

semantic landmarks

$$\hat{\mathbf{d}}(w_r) = \theta\,\mathbf{d}(w_r^0) + \frac{1}{\mathcal{L}_r}\sum_{i=1}^{|\mathcal{L}_r|} \mathrm{e}^{-i}\mathbf{d}(l_{i,r})$$

induced embedding for w_r

embedding for the i-th landmark

if there exists an initial (unreliable) representation for w_r which is to be improved

Inducing embedding for unseen words

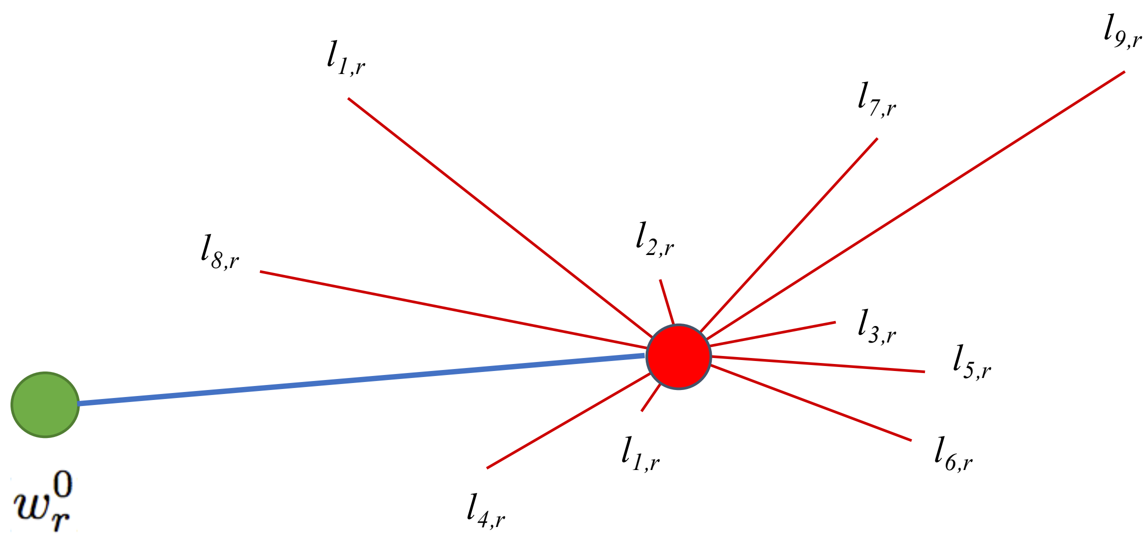
3. Induce embeddings

semantic landmarks

$$\hat{\mathbf{d}}(w_r) = \theta\, \mathbf{d}(w_r^0) + \frac{1}{\mathcal{L}_r} \sum_{i=1}^{|\mathcal{L}_r|} \mathrm{e}^{-i} \mathbf{d}(l_{i,r})$$

induced embedding for w_r

embedding for the i-th landmark



Experiments

General domain setting

Rare Word similarity dataset

External lexical resource: WordNet

	Vanilla			+Induction		
	OOV	r	ρ	OOV	r	ρ
GLOVE	11%	34.9	34.4	0%	38.6	39.7
W2V-250K	34%	31.0	25.9	0%	44.2	47.5
W2V-GN	9%	43.8	45.3	0%	48.3	50.5

Experiments

General domain setting

Rare Word similarity dataset

Approach	RW		RG-65	
	OOV	ρ	OOV	ρ
Botha and Blunsom (2014)	NA	30.0	NA	41.0
Luong et al. (2013)*	0%	34.4	0%	65.5
Soricut and Och (2015)*	0%	41.8	0%	75.1
Our approach*	0%	**43.3**	0%	75.1
Number of pairs	2034		65	

Systems marked with * are trained on the same corpus.

Experiments

Specific domain setting (medical)

Datasets: MayoSRS (101 pairs) and UMNSRS (566 pairs)

External lexical resource: Medical Subject Headings (MeSH)

		Vanilla			**+Induction**		
		OOV	r	ρ	OOV	r	ρ
Mayo	GloVe	16%	11.1	11.6	11%	36.7	26.1
	w2v-250k	41%	1.2	2.9	21%	27.8	20.1
	w2v-GN	12%	15.5	14.0	10%	18.4	10.9
UMN	GloVe	17%	31.6	24.4	6%	38.2	33.6
	w2v-250k	38%	11.8	3.2	13%	27.8	20.1
	w2v-GN	17%	25.8	21.5	7%	32.8	32.4

Conclusions

- A novel approach to inducing embeddings for unseen words

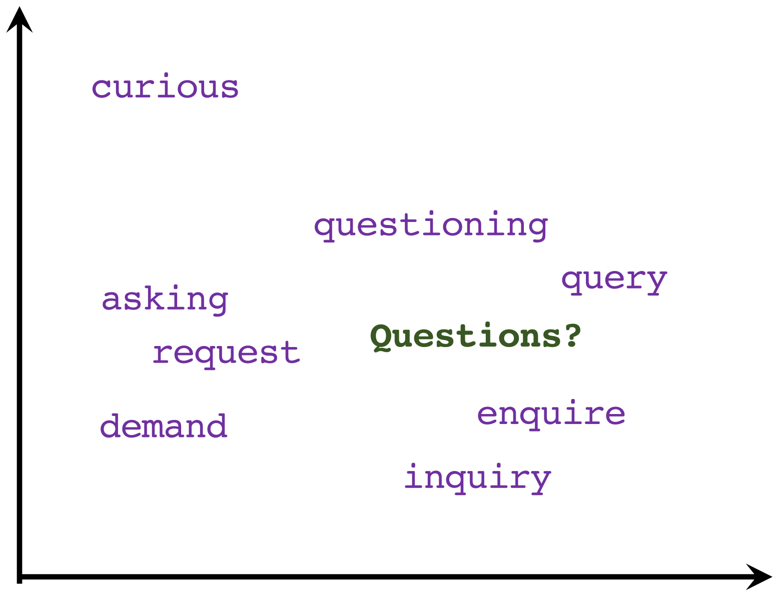
  Based on the knowledge encoded in external lexical resources

- Improved performance on multiple benchmarks in general and specific domains

❖ Extension to other domains and languages

❖ New evaluation benchmarks

Thank you!


curious
questioning
query
asking
Questions?
request
enquire
demand
inquiry