Probing Taxonomic and Thematic Embeddings for Taxonomic Information

Filip Klubička, John D. Kelleher

24th January 2023

Global WordNet Conference, Donostia-San Sebastian

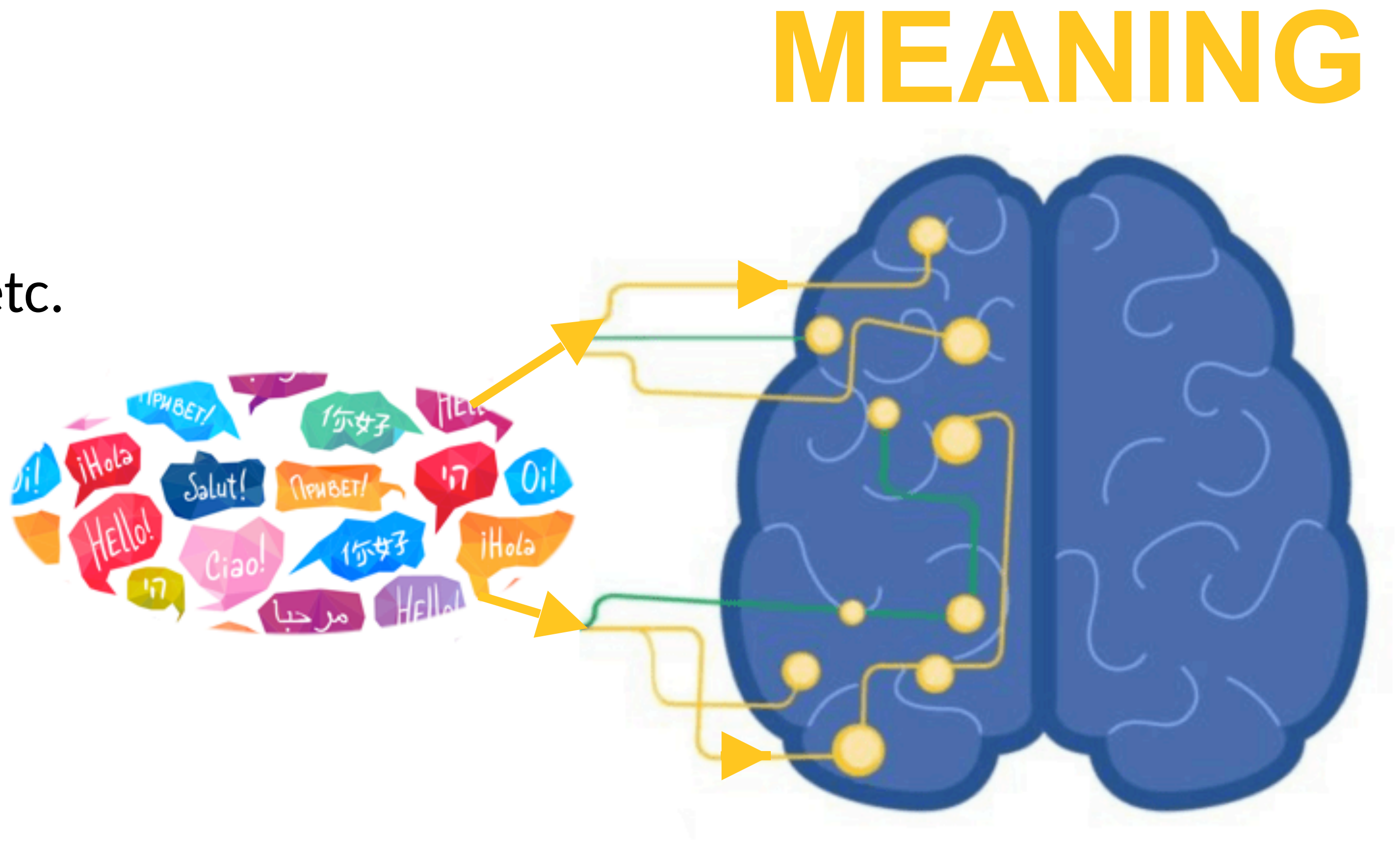
I. Introduction

Engaging Content
Engaging People

Semantics

Engaging Content
Engaging People

- semantics
- meaning
- semantic relations
- taxonomies, ontologies etc.



Natural Language Processing

- computational semantics
- meaning representations
- vector space models
- embeddings (word2vec, GLOVE...)
- language models (BERT, GPT-3...)

Interpretability

- explainable AI
- interpretable models
- BlackboxNLP (Alishahi et al. 2019)
- probing framework

II. Background and Motivation

Engaging Content
Engaging People

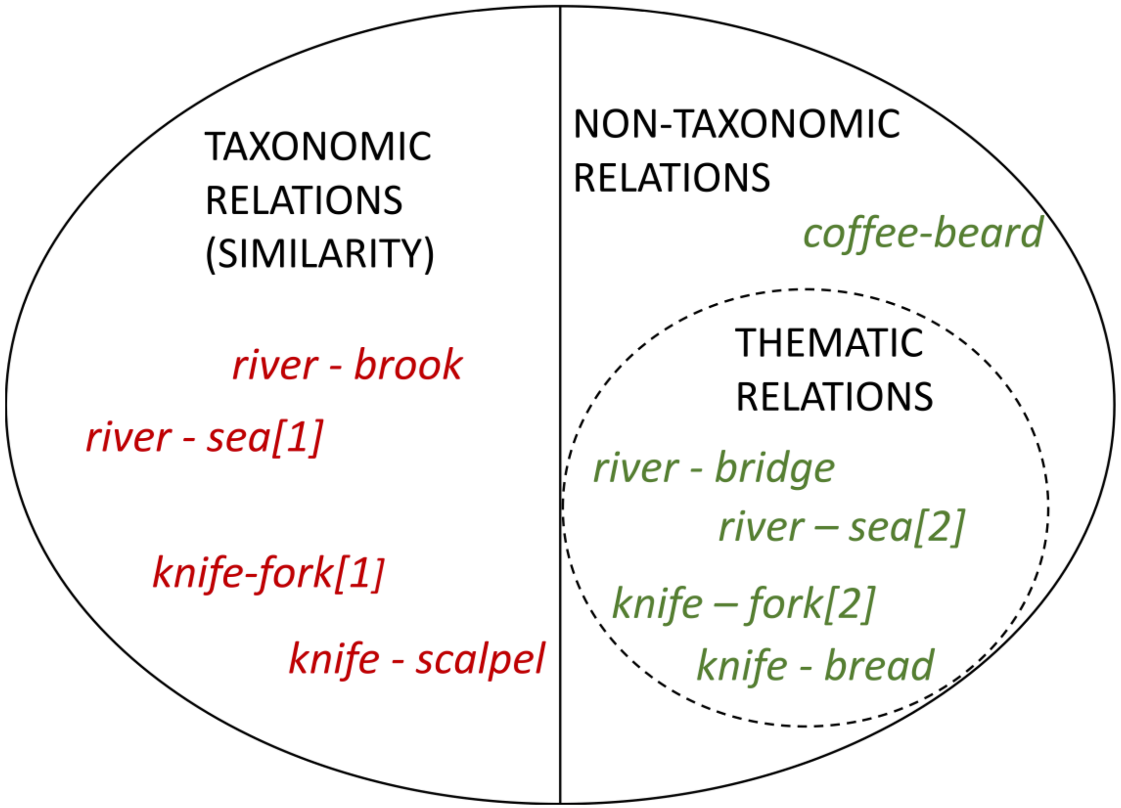
Semantic similarity

- semantic similarity encompasses a variety of lexico-semantic and topical relations
- distributional semantics literature often underspecifies what kind of similarity is being modeled (Kacmajor and Kelleher, 2019)

Two key dimensions of semantic relationships

- **taxonomic**
- **non-taxonomic**

Types of semantic relatedness



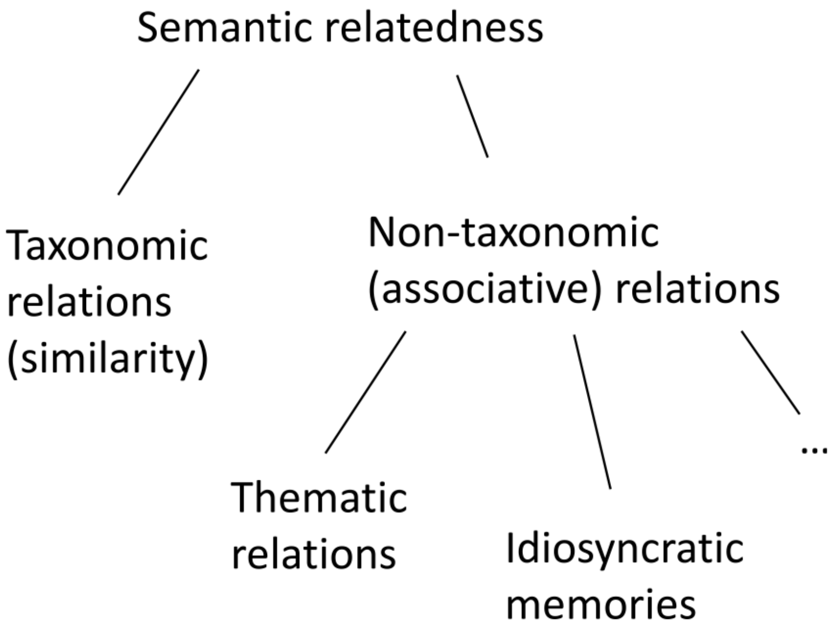


Figure 1. Subsets of semantic relatedness. (Image originally published by Kacmajor and Kelleher (2019) as Figure 1, licensed under CC BY 4.0.)

Types of semantic relatedness

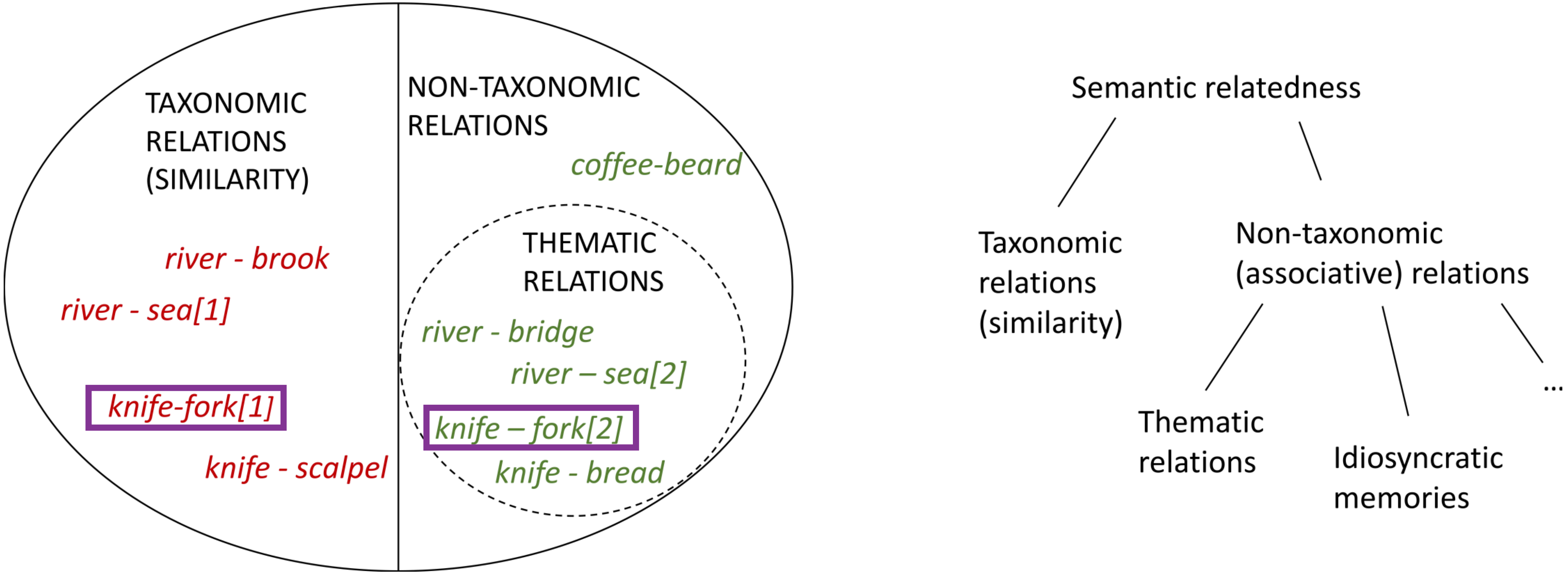


Figure 1. Subsets of semantic relatedness. (Image originally published by Kacmajor and Kelleher (2019) as Figure 1, licensed under CC BY 4.0.)

Conflation of semantic relationships

- in the distributional semantics literature *taxonomic* and *thematic* similarity is often conflated (Kacmajor and Kelleher, 2019)
- the word *similarity* most often refers to *taxonomic similarity*
  - this is usually not explicitly stated
- an important distinction; differentiating could improve statistical language models
  - taxonomic relations indicate replacement
  - thematic relations help in predicting the next word in a sequence

Linguistic resources

- different language resources reflect different semantic relationships

Knowledge-Engineered Resources

- thesauri, knowledge bases, ontologies, taxonomies, semantic networks
- explicitly encode and reflect *taxonomic relations*

Natural Language Corpora

- only provide linguistic context and word co-occurrence information
- encode and reflect *thematic relations*

- if a language model is trained on just one type of resource, arguably it cannot accurately encode the full spectrum of semantic relatedness

Research questions

- How much taxonomic information is encoded in thematic embeddings?
- How much taxonomic information is encoded in taxonomic embeddings?
- Are there differences in how this information is encoded vector space?

Approach:

- apply the probing framework
- develop taxonomic probing dataset based on English WordNet
- examine differences in structural properties of taxonomic and thematic embedding space

III. Probing Classifiers

Engaging Content
Engaging People

Probing classifiers

- in essence a linguistic classification task
- uses “vanilla” language embeddings as input to ML classifier (probe)
- probe predicts some linguistic property of interest
  - e.g. sentence length, verb tense, subject number, parse tree depth etc.
  - particularly interesting to examine linguistic properties which the models are not explicitly trained to encode, thus revealing emergent structures
- **intuition**: if the probe performs well, the relevant knowledge must be encoded in the representation

Probing at a glance

1. Choose a linguistic property of interest, e.g. verb tense
2. Choose or design an appropriate dataset
3. Choose a word/sentence representation, e.g. BERT
4. Choose a probing classifier (i.e. the probe), e.g. MLP
5. Train the probe on the embeddings as input
6. Evaluate the probe’s performance on the task

Some challenges

How to determine if the probe performs well?

- probe interpretations are inherently comparative
- goal: move towards "intrinsic" probe evaluations

Focus on vector dimensions—what about the norm?

- norm is rarely studied and often overlooked (e.g. cosine similarity normalises vectors)
- goal: exploration of the role of the norm in encoding information

Information containers

Embeddings = Vectors

- vectors = direction + magnitude
- direction (coordinates) defined by dimension values
- magnitude (length) defined by vector norm

vector									norm
10	5	-2	4	-8	1	2	5		37

- two information containers
  - vector **dimensions**
  - vector **norm**

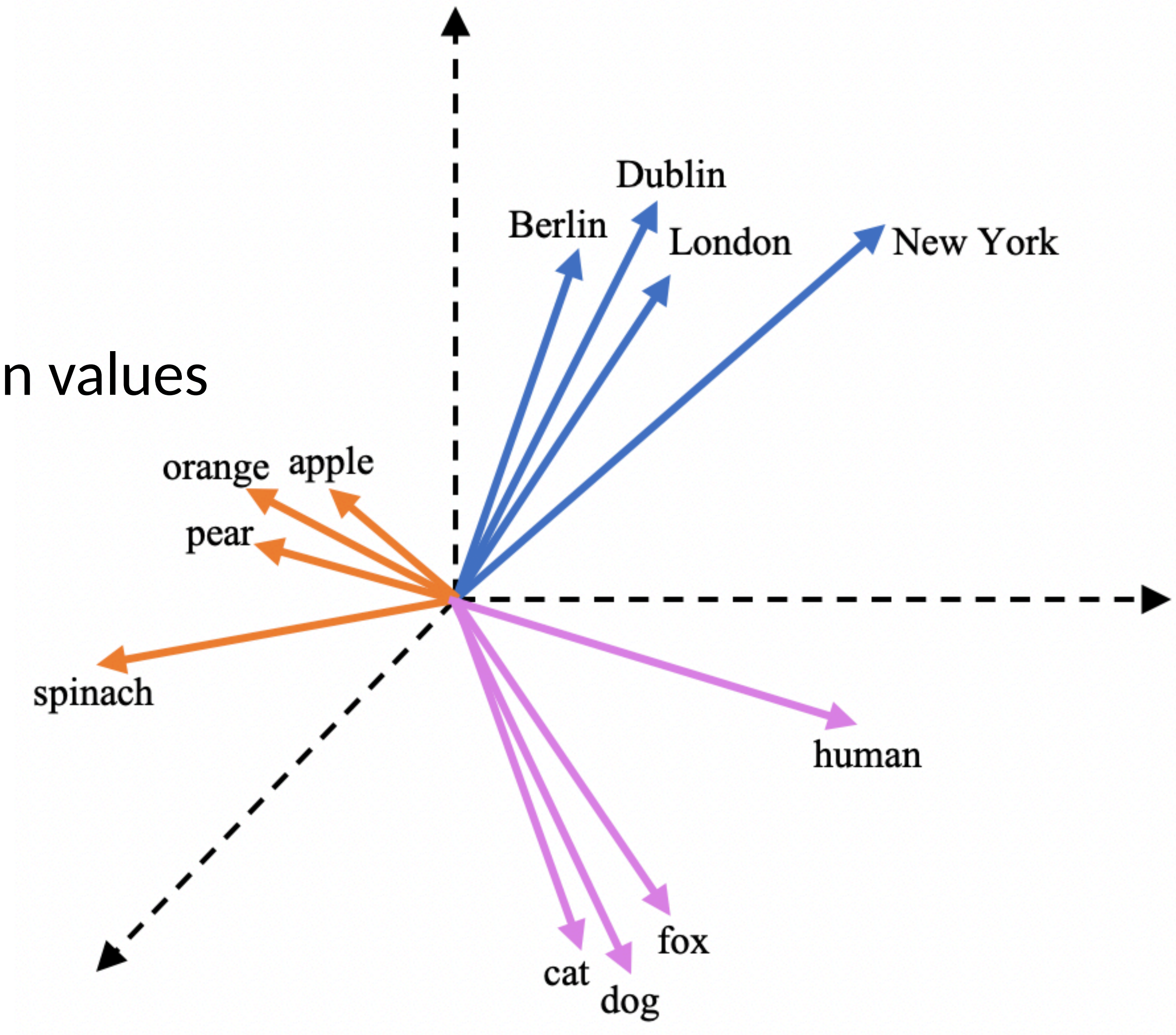


Figure 2. An illustrative example of a vector space model.

Probing with Noise

1. Choose a linguistic property of interest, e.g. verb tense
2. Choose or design an appropriate dataset
3. Choose a word/sentence representation, e.g. BERT
4. Choose a probing classifier (i.e. the probe), e.g. MLP
5. Train the probe on the embeddings as input
6. Evaluate the probe’s performance on the task

Probing with Noise

1. Choose a linguistic property of interest, e.g. verb tense
2. Choose or design an appropriate dataset
3. Choose a word/sentence representation, e.g. BERT
4. Choose a probing classifier (i.e. the probe), e.g. MLP
5. Train the probe on the embeddings as input
6. Evaluate the probe's performance on the task (vanilla baseline)
7. Introduce systematic noise in the embedding
8. Repeat training, evaluate and compare

IV. Taxonomic Probing Dataset

Engaging Content
Engaging People

Motivation

- a probing task needs to ask a simple, non-ambiguous question
- hypernym detection/discovery and cloze tasks not ideally suited to our framework
- require a simpler task that more directly teases out the hypernym-hyponym relationship
- new taxonomic probing task: predicting which word in a pair is the hypernym, and which is the hyponym
  - derived from WordNet
  - each pair shares an immediate hypernym-hyponym relationship
  - a word in a pair can **only** be a direct hyponym or hypernym of the other

Dataset construction

- dataset pruning: only contains the intersection of vocabularies of our encoders
  - only includes word pairs that have representations in all our embedding models
- problem definition: positional classification task
  - concatenate word vectors in the pair
  - **Q**: given a pair of words, is the first one the hyponym (0) or hypernym (1) of the other?
- balancing: duplicate all instances and swap the positions in the pair

- Final set: 493,494 word pairs, 50,000 in test set, remainder in training set
  - **0**, north, direction
  - **1**, direction, north
  - **0**, hurt, upset
  - **1**, upset, hurt

V. Probing Experiments

Engaging Content
Engaging People

Models

Thematic Embeddings

- **SGNS**
  - genism word2vec implementation
  - Google News dataset
  - 300-dimensional word embedding
  - off-the-shelf
- **GloVe**
  - common crawl (2.2M tokens), cased
  - 300-dimensional word embedding
  - off-the-shelf

Taxonomic Embeddings

- **SGNS**
  - taxonomic WordNet random walk embeddings (Klubička et al., 2019)
  - 300-dimensional word embedding
  - off-the-shelf
- **GloVe**
  - trained on same taxonomic pseudo-corpora as SGNS above (Klubička et al. 2020)
  - 300-dimensional word embedding

- final instances in the dataset contain 2 concatenated vectors = 600 dimensions

Evaluation

- probe model: Multi-Layered Perceptron (MLP)
- evaluation metric: AUC_ROC score (0.5 = model does not discriminate)
- train 50 times and report average scores

Questions:

- How do *vanilla embeddings* perform on the task ?
- What is the effect of *ablated norm* vs *ablated dimensions* ?

Results SGNS – THEM vs TAX

	SGNS			
Model	THEM		TAX	
	auc	±CI	auc	±CI
rand. pred.	.5000	.0009	.4997	.0009
rand. vec.	.5001	.0012	.5001	.0011
vanilla	.9163	.0004	.9256	.0003

Table 1. Evaluation scores of the probing with noise experiments on taxonomic and thematic SGNS embeddings. Reporting AUC_ROC evaluation scores and the confidence interval (CI) of the average calculated over all training runs.

Results SGNS – THEM vs TAX

	SGNS			
Model	THEM		TAX	
	auc	±CI	auc	±CI
rand. pred.	.5000	.0009	.4997	.0009
rand. vec.	.5001	.0012	.5001	.0011
vanilla	.9163	.0004	.9256	.0003
abl. N	.9057	.0004	.9067	.0005
abl. D	.5039	.0008	.5294	.0010
abl. D+N	.4998	.0010	.5002	.0009

Table 1. Evaluation scores of the probing with noise experiments on taxonomic and thematic SGNS embeddings. Reporting AUC_ROC evaluation scores and the confidence interval (CI) of the average calculated over all training runs.

Results SGNS – THEM vs TAX

	SGNS			
Model	THEM		TAX	
	auc	±CI	auc	±CI
rand. pred.	.5000	.0009	.4997	.0009
rand. vec.	.5001	.0012	.5001	.0011
vanilla	.9163	.0004	.9256	.0003
abl. N	.9057	.0004	.9067	.0005
abl. D	.5039	.0008	.5294	.0010
abl. D+N	.4998	.0010	.5002	.0009

Table 1. Evaluation scores of the probing with noise experiments on taxonomic and thematic SGNS embeddings. Reporting AUC_ROC evaluation scores and the confidence interval (CI) of the average calculated over all training runs.

Results SGNS – THEM vs TAX

	SGNS			
Model	THEM		TAX	
	auc	±CI	auc	±CI
rand. pred.	.5000	.0009	.4997	.0009
rand. vec.	.5001	.0012	.5001	.0011
vanilla	.9163	.0004	.9256	.0003
abl. N	.9057	.0004	.9067	.0005
abl. D	.5039	.0008	.5294	.0010
abl. D+N	.4998	.0010	.5002	.0009

Table 1. Evaluation scores of the probing with noise experiments on taxonomic and thematic SGNS embeddings. Reporting AUC_ROC evaluation scores and the confidence interval (CI) of the average calculated over all training runs.

Results SGNS – THEM vs TAX

SGNS				
Model	THEM		TAX	
	auc	±CI	auc	±CI
rand. pred.	.5000	.0009	.4997	.0009
rand. vec.	.5001	.0012	.5001	.0011
vanilla	.9163	.0004	.9256	.0003
abl. N	.9057	.0004	.9067	.0005
abl. D	.5039	.0008	.5294	.0010
abl. D+N	.4998	.0010	.5002	.0009

Table 1. Evaluation scores of the probing with noise experiments on taxonomic and thematic SGNS embeddings. Reporting AUC_ROC evaluation scores and the confidence interval (CI) of the average calculated over all training runs.

Results SGNS – THEM vs TAX

	SGNS			
Model	THEM		TAX	
	auc	±CI	auc	±CI
rand. pred.	.5000	.0009	.4997	.0009
rand. vec.	.5001	.0012	.5001	.0011
vanilla	.9163	.0004	$\boxed{.9256}$	.0003
abl. N	.9057	.0004	.9067	.0005
abl. D	.5039	.0008	$\boxed{.5294}$	.0010
abl. D+N	.4998	.0010	.5002	.0009

Table 1. Evaluation scores of the probing with noise experiments on taxonomic and thematic SGNS embeddings. Reporting AUC_ROC evaluation scores and the confidence interval (CI) of the average calculated over all training runs.

Results SGNS – THEM vs TAX

	SGNS			
Model	THEM		TAX	
	auc	±CI	auc	±CI
rand. pred.	.5000	.0009	.4997	.0009
rand. vec.	.5001	.0012	.5001	.0011
vanilla	.9163	.0004	.9256	.0003
abl. N	.9057	.0004	.9067	.0005
abl. D	.5039	.0008	.5294	.0010
abl. D+N	.4998	.0010	.5002	.0009

Table 1. Evaluation scores of the probing with noise experiments on taxonomic and thematic SGNS embeddings. Reporting AUC_ROC evaluation scores and the confidence interval (CI) of the average calculated over all training runs.

Results GloVe – THEM vs TAX

	GloVe			
Model	THEM		TAX	
	auc	±CI	auc	±CI
rand. pred.	.4999	.0011	.4998	.0010
rand. vec.	.5001	.0010	.5001	.0008
vanilla	.9327	.0004	.8824	.0005
abl. N	.9110	.0004	.8435	.0008
abl. D	.5002	.0008	$\boxed{.6621}$	.0008
abl. D+N	.5000	.0011	.5006	.0011

Table 2. Evaluation scores of the probing with noise experiments on taxonomic and thematic GloVe embeddings. Reporting AUC_ROC evaluation scores and the confidence interval (CI) of the average calculated over all training runs.

Results GloVe – THEM vs TAX

	GloVe			
Model	THEM		TAX	
	auc	±CI	auc	±CI
rand. pred.	.4999	.0011	.4998	.0010
rand. vec.	.5001	.0010	.5001	.0008
vanilla	.9327	.0004	.8824	.0005
abl. N	.9110	.0004	.8435	.0008
abl. D	.5002	.0008	.6621	.0008
abl. D+N	.5000	.0011	.5006	.0011

Table 2. Evaluation scores of the probing with noise experiments on taxonomic and thematic GloVe embeddings. Reporting AUC_ROC evaluation scores and the confidence interval (CI) of the average calculated over all training runs.

Findings: vanilla embeddings

- both taxonomic and thematic SGNS and GloVe encode some taxonomic information
- taxonomic SGNS encodes significantly more taxonomic information than thematic SGNS
- thematic GloVe encodes the most taxonomic information compared to other embeddings

Structural insights

Engaging Content
Engaging People

- the norm can carry linguistic information at the *word level*
- different encoders utilise the norm to *varying degrees*
  - taxonomic GloVe encodes *more* taxonomic information in the norm than word2vec
  - thematic GloVe encodes *no* taxonomic information in its norm
- *taxonomic embeddings* encode more taxonomic information in the norm than thematic embeddings to
  - the norm is used to supplement encoding of taxonomic information
- the usage of the norm can be determined by the embedding training data, i.e. the *underlying distribution*, rather than the model architecture

VI. Additional Analyses

Engaging Content
Engaging People

Hypernym-hyponym norm lengths - SGNS

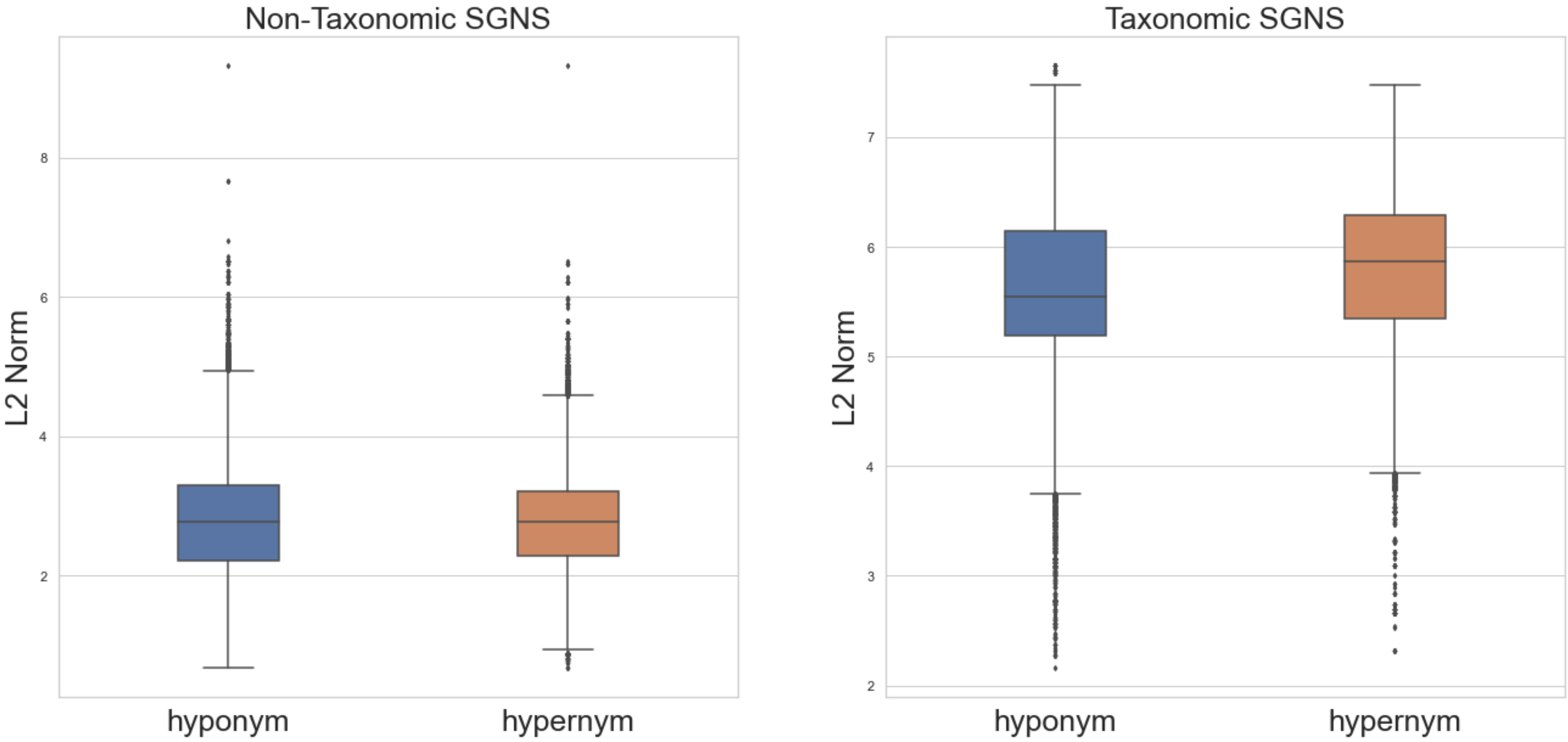


Figure 3. Box plots depicting the median values of the L2 norm in the different sets of word vectors, separate for hyponyms and hypernyms. There is a marked difference observed between hyponym and hypernym norms in taxonomic GloVe and SGNS, but not in thematic.

Hypernym-hyponym norm lengths - GloVE

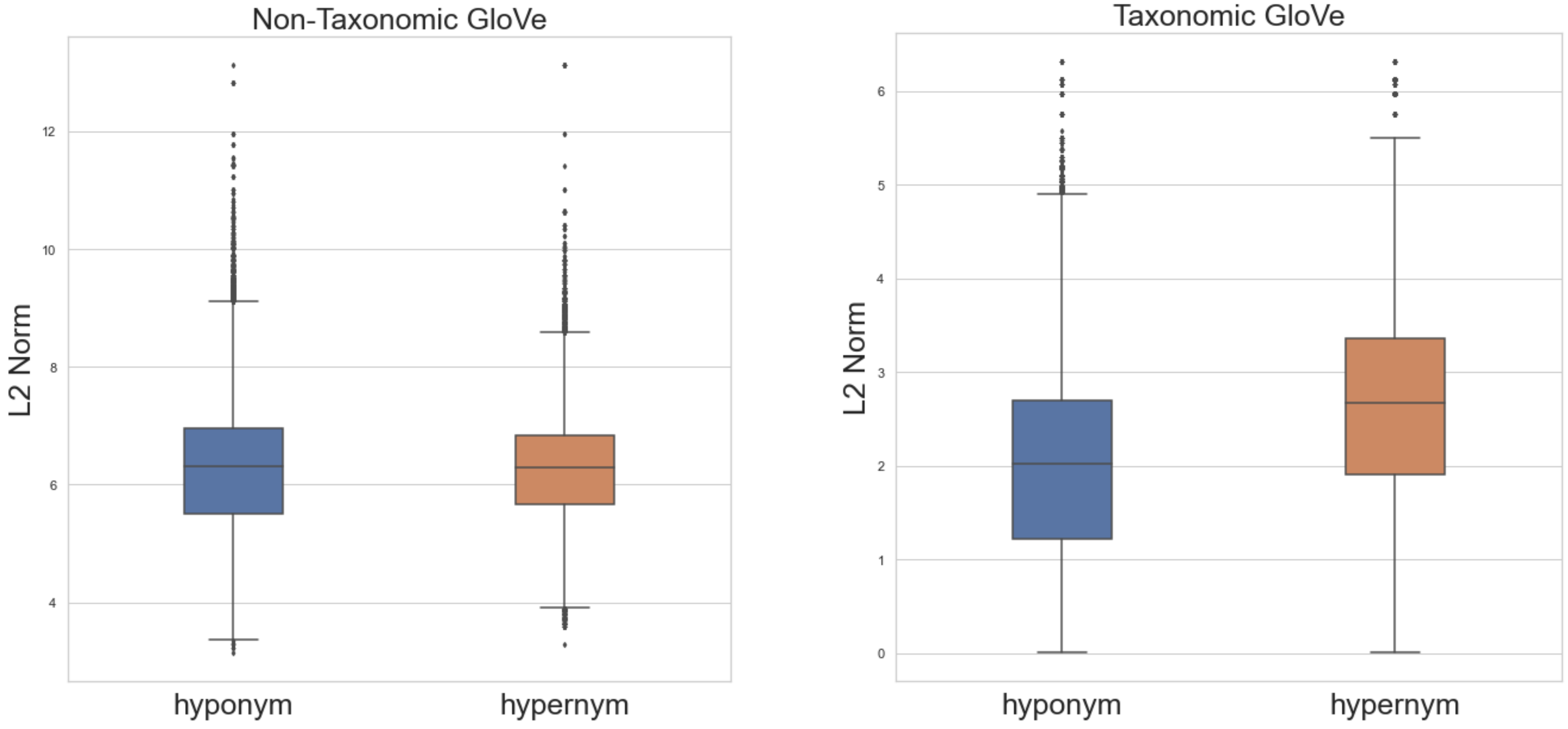


Figure 4. Box plots depicting the median values of the L2 norm in the different sets of word vectors, separate for hyponyms and hypernyms. There is a marked difference observed between hyponym and hypernym norms in taxonomic GloVe and SGNS, but not in thematic.

Norm lengths observation

- on average, the norm of hypernyms is *longer* than the norm of hyponyms
  - only in taxonomic embeddings
- there is a mapping between the taxonomic hierarchy and distance from the origin
  - *hypernyms* (higher in taxonomy) are *further away from the origin*
  - *hyponyms* (lower in taxonomy) and are *closer to the origin*

VII. Conclusion

Engaging Content
Engaging People

Recap

- applied *probing with noise* to *taxonomic* and *thematic SGNS* and *GloVe* embeddings
- designed new *taxonomic probing task* derived from WordNet
- *both* taxonomic and thematic embeddings encode taxonomic information
  - taxonomic SGNS embeddings encode *more*
- the probe is using the *relationship* between candidate words as a *predictive feature*
- provide *geometric insight* into the vector space and role of the norm in encoding taxonomic information
  - GloVe encodes *a lot* of taxonomic information in the norm
  - taxonomic embeddings use the norm to supplement their encoding of taxonomic information
- the usage of the norm can be determined by the embedding training data, i.e. the *underlying distribution*, rather than the model architecture

Thank you for your attention!

e-mail: filip.klubicka@adaptcentre.ie

twitter: @lemoncloak

github: https://github.com/GreenParachute

School of Computer Science

Technological University Dublin

www.adaptcentre.ie

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH

TU DUBLIN

TECHNOLOGICAL UNIVERSITY DUBLIN

ADAPT

Engaging Content
Engaging People

FUNDED BY

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union

European Union
European Regional Development Fund

Science Foundation Ireland sfi For what's next

Taxonomic relatedness

- relationship based on a comparison of the concepts' features
- taxonomically related words/concepts share properties or functions

- *table vs. desk*

Non-taxonomic relatedness

- related by virtue of co-occurrence in any sort of context (e.g. temporal, spatial, linguistic)

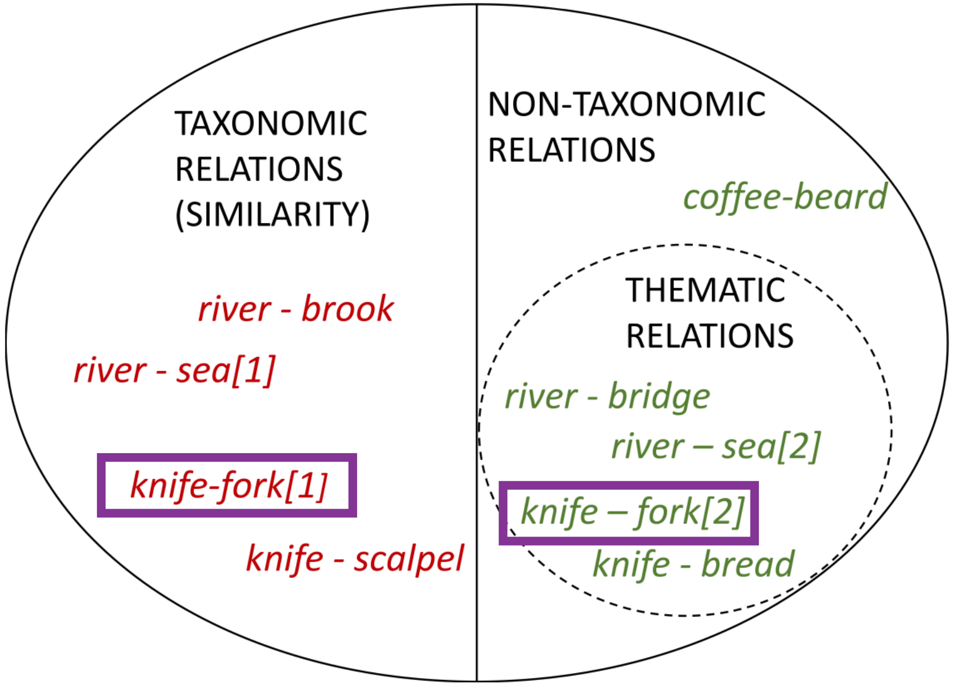
Thematic relations (Lin and Murphy, 2001)

- thematically related words/concepts perform complementary roles in a common event or theme
- this often implies having different features and functions which are complementary
- *table* vs. *chair*

- distributionally, thematic relations reflect high-probability co-occurrences

¿Por qué no los dos?

- Kacmajor and Kelleher (2019) show that the same pair of concepts can be connected via both taxonomic and thematic relations



Paradigmatic vs. syntagmatic relationships

- taxonomic & thematic **≈** paradigmatic & syntagmatic (De Saussure, 1916)

The Sun is shining.

Paradigmatic

- *vertical*
- relationship among linguistic elements that can substitute for each other in a given context
- *Sun* ⇄ *Moon* ⇄ *stars* ⇄ *light*

Syntagmatic

- *horizontal*
- relationship among linguistic elements that occur sequentially in a chain of speech/text
- *The Sun* ⇄ *is shining*

- *substitution* vs. *positioning*

The method’s supporting pillars

a) systematic noise – helps ablate information
b) random baselines – basis for relative intrinsic evaluation
c) confidence intervals – inform inferences

A) Systematic noise

- ablating a vector's information containers individually
- noise should not affect both containers

- solution: random sampling + scaling
  - dimension container: random dimension values scaled to original norm
  - norm container: existing dimension values scaled to random norm

- both containers can be affected by introducing both types of noise at the same time: this can act as a sense check

B) Random baselines

- grounding the impact of vector modifications
- problem: probe can learn class distributions
- baselines:
  a) random prediction on test set
  b) train probe on randomly generated vectors

C) Confidence intervals

Randomness in probing with noise

- probe might contain a stochastic component
- noising functions are highly stochastic
- evaluation scores will vary when probe is retrained

Solution

- train model a multitude of times and report average score
- 99% confidence interval provides statistical significance
- confidence interval range used when comparing models

Dimension deletion experiments

SGNS				
Model	THEM		TAX	
	auc	±CI	auc	±CI
rand. pred.	.5000	.0009	.4997	.0009
rand. vec.	.5001	.0012	.5001	.0011
vanilla	.9163	.0004	.9256	.0003
del. ea. 1h	.8929	.0004	.8998*	.0005
del. ea. 2h	.8927	.0004	.9039	.0004
del. ct. 1h	.8496	.0004	.8525	.0004
del. ct. 2h	.8495	.0004	.8523	.0003

Table 3. Probing results on SGNS deletions and baselines. Reporting average AUC-ROC scores and confidence intervals (CI) of the average of all training runs

Dimension deletion experiments

	SGNS			
Model	THEM		TAX	
	auc	±CI	auc	±CI
rand. pred.	.5000	.0009	.4997	.0009
rand. vec.	.5001	.0012	.5001	.0011
vanilla	.9163	.0004	.9256	.0003
del. ea. 1h	.8929	.0004	.8998*	.0005
del. ea. 2h	.8927	.0004	.9039	.0004
del. ct. 1h	.8496	.0004	.8525	.0004
del. ct. 2h	.8495	.0004	.8523	.0003

Table 3. Probing results on SGNS deletions and baselines. Reporting average AUC-ROC scores and confidence intervals (CI) of the average of all training runs

Dimension deletion experiments

	SGNS			
Model	THEM		TAX	
	auc	±CI	auc	±CI
rand. pred.	.5000	.0009	.4997	.0009
rand. vec.	.5001	.0012	.5001	.0011
vanilla	.9163	.0004	.9256	.0003
del. ea. 1h	.8929	.0004	.8998*	.0005
del. ea. 2h	.8927	.0004	.9039	.0004
del. ct. 1h	.8496	.0004	.8525	.0004
del. ct. 2h	.8495	.0004	.8523	.0003

Table 3. Probing results on SGNS deletions and baselines. Reporting average AUC-ROC scores and confidence intervals (CI) of the average of all training runs

Engaging Content
Engaging People

Dimension deletion experiments

	GloVe			
Model	THEM		TAX	
	auc	±CI	auc	±CI
rand. pred.	.4999	.0011	.4998	.0010
rand. vec.	.5001	.0010	.5001	.0008
vanilla	.9327	.0004	.8824	.0005
del. ea. 1h	.9120*	.0003	.8727	.0005
del. ea. 2h	.9179	.0004	.8730	.0006
del. ct. 1h	.8522	.0004	.8405	.0004
del. ct. 2h	.8522	.0004	.8406	.0004

Table 4. Probing results on SGNS deletions and baselines. Reporting average AUC-ROC scores and confidence intervals (CI) of the average of all training runs

Dimension deletion experiments

	GloVe			
Model	**THEM**		**TAX**	
	auc	±CI	auc	±CI
rand. pred.	.4999	.0011	.4998	.0010
rand. vec.	.5001	.0010	.5001	.0008
vanilla	.9327	.0004	.8824	.0005
del. ea. 1h	.9120*	.0003	.8727	.0005
del. ea. 2h	.9179	.0004	.8730	.0006
del. ct. 1h	.8522	.0004	.8405	.0004
del. ct. 2h	.8522	.0004	.8406	.0004

Table 4. Probing results on SGNS deletions and baselines. Reporting average AUC-ROC scores and confidence intervals (CI) of the average of all training runs

Deletion findings

- larger drop in both *del.ct.* settings versus *del.ea.* settings
  - predicting a word's relationship to an “imaginary” other word is the *more difficult* task
- in both cases performance *significantly above random*
  - probe learned some frequency distributions from the graph
  - reflects hypernym-hyponym imbalance inherent to WordNet
- learning from two halved vectors is better than a single full representation
  - probe is *inferring the relevant relationship* between the candidate words