Analysis of Word Embeddings: A Clustering and Topological Approach

Master's thesis defence, 30 June 2021

By: Jonas Folkvord Triki
Supervisor: Nello Blaser

UNIVERSITY OF BERGEN

Agenda

1. Introduction
2. Methods
3. Results
4. Conclusion
5. Future Work

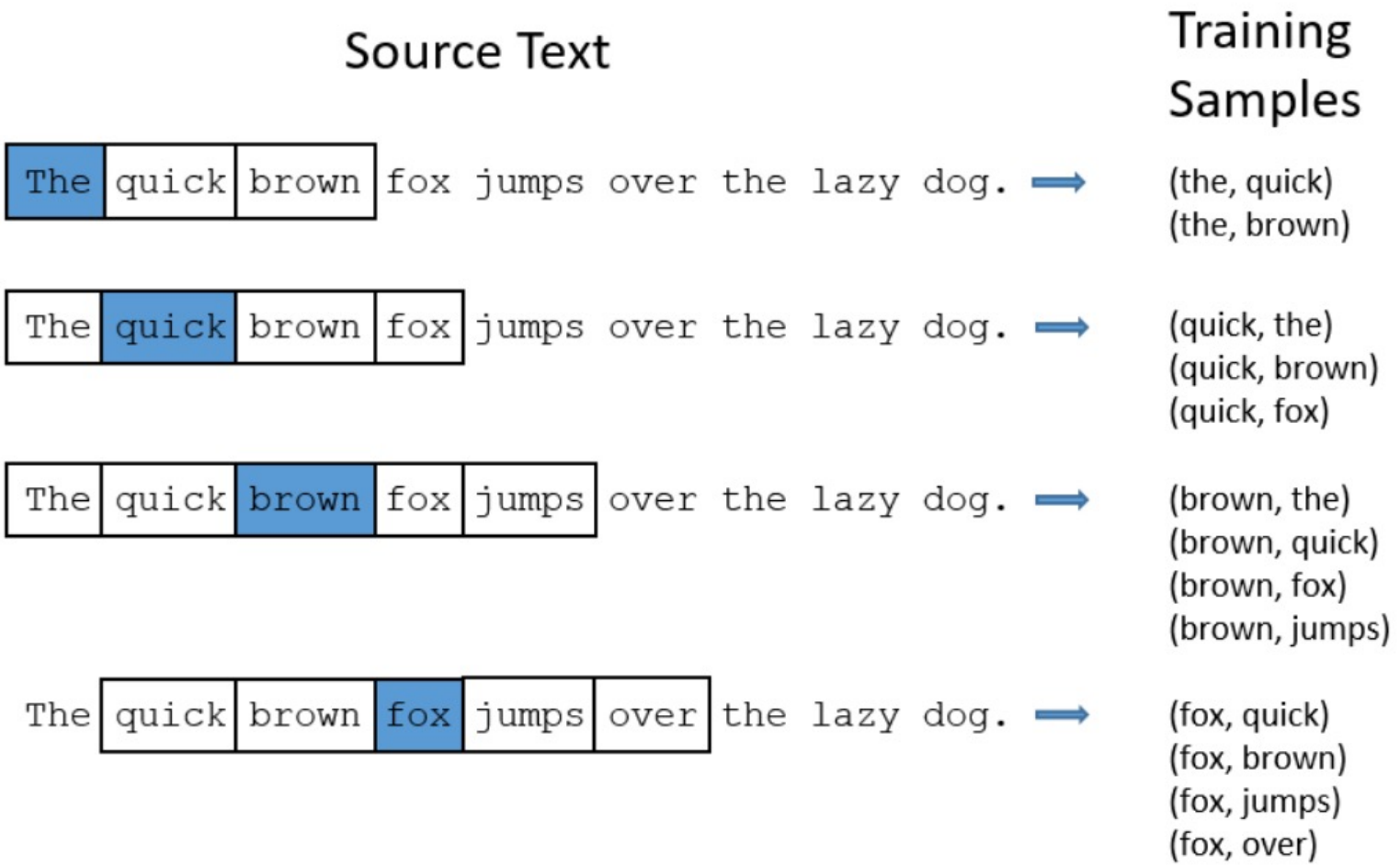
Introduction

Introduction

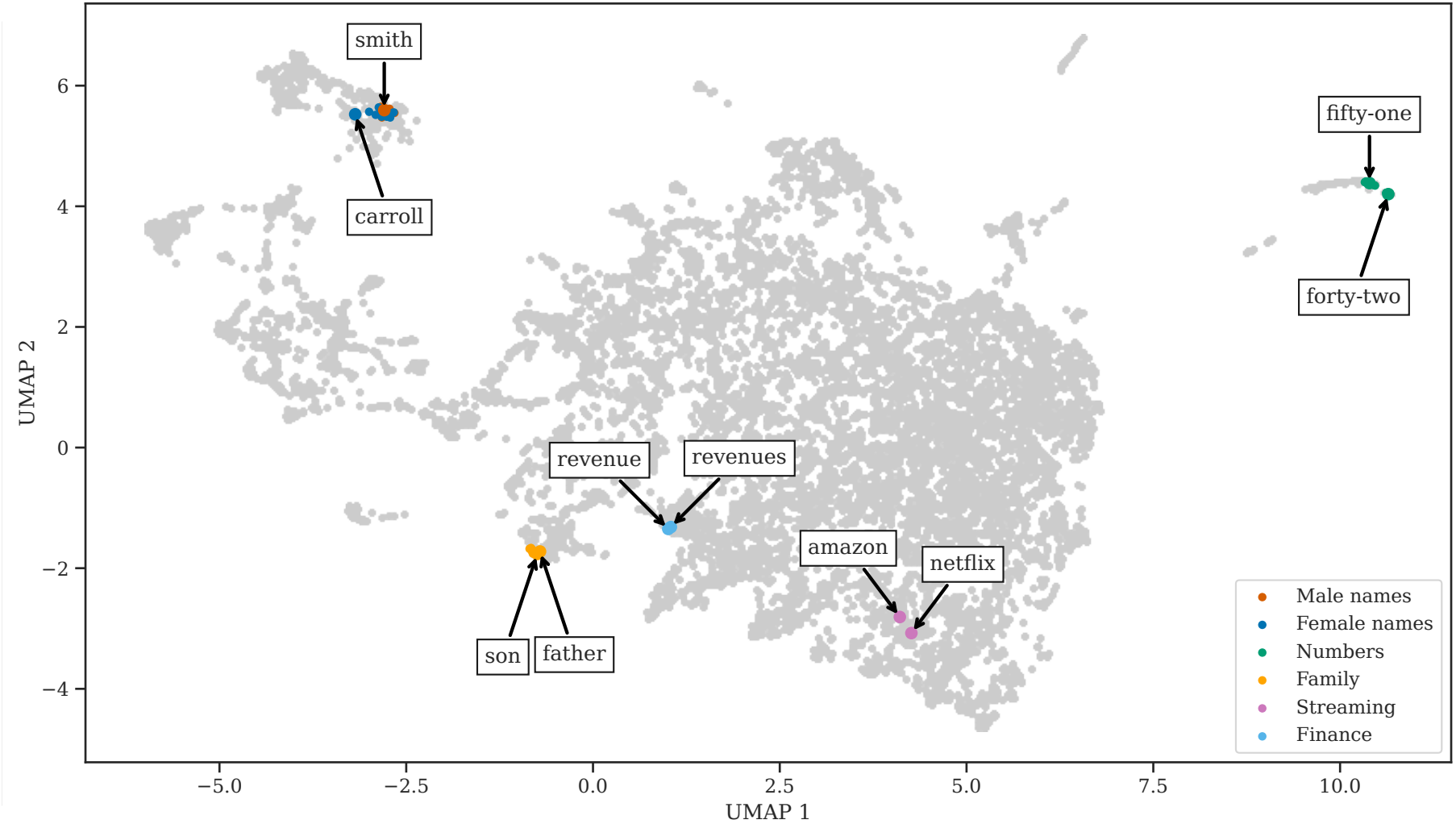
- Natural language processing (NLP)
- Words to word embeddings
- Single vector representations
- Polysemy of words, e.g.
  - run
  - open
  - make

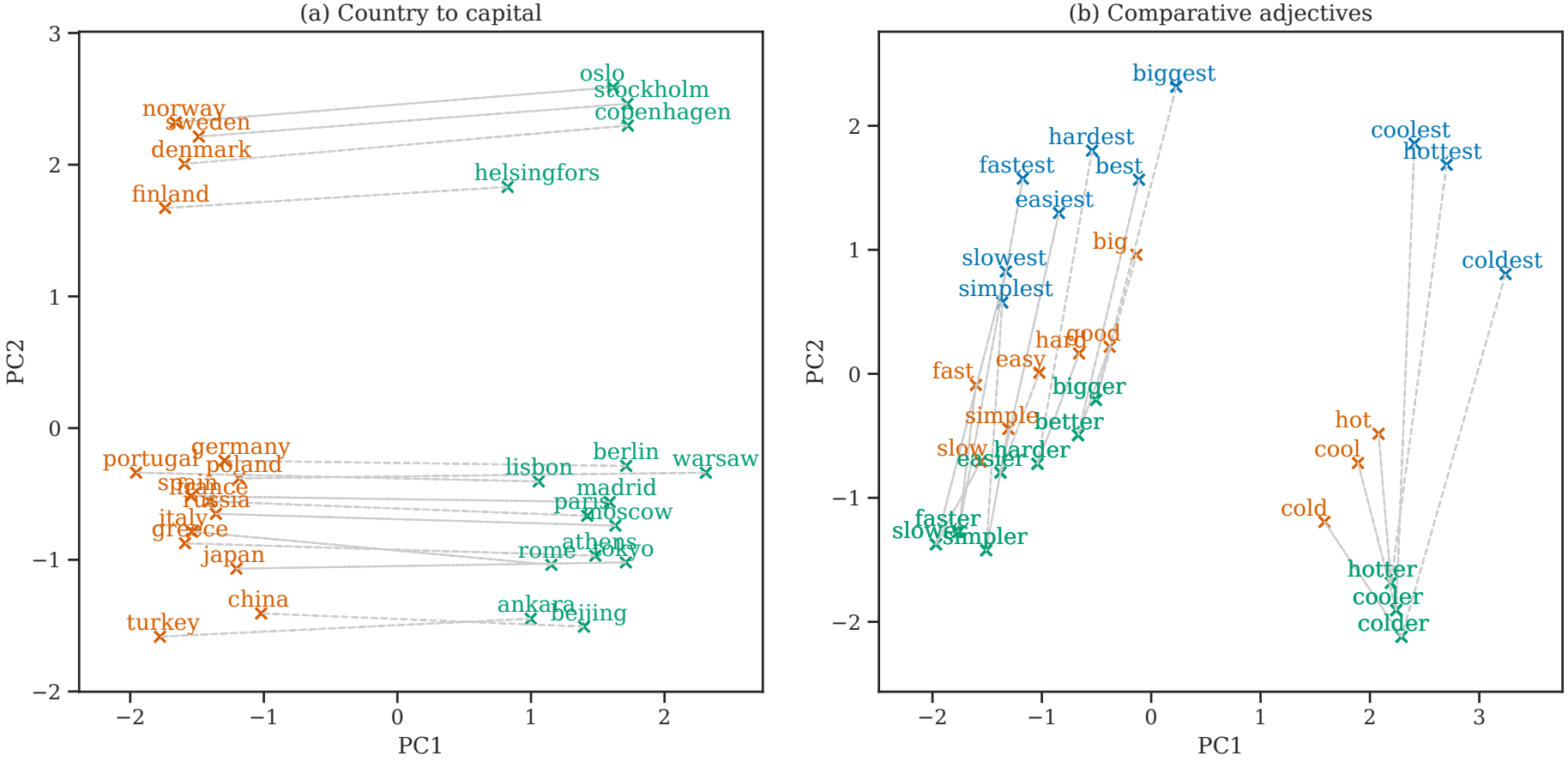
Word embeddings

- Vector representations of words
  - Similar words have similar vector representations
  - 300 dimensions
- Word embedding models, e.g.
  - word2vec
  - GloVe
  - fastText



Source: https://medium.com/@zafaralibagh6/simple-tutorial-on-word-embedding-and-word2vec-43d477624b6d, downloaded 13 June 2021.


smith
carroll
fifty-one
forty-two
revenue
revenues
amazon
netflix
son
father
Male names
Female names
Numbers
Family
Streaming
Finance
UMAP 1
UMAP 2
6
4
2
0
−2
−4
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0


(a) Country to capital
PC1
PC2
norway
sweden
denmark
finland
oslo
stockholm
copenhagen
helsingfors
portugal
germany
poland
spain
france
russia
italy
greece
japan
china
turkey
berlin
warsaw
lisbon
madrid
paris
moscow
athens
rome
tokyo
ankara
beijing
(b) Comparative adjectives
PC1
PC2
biggest
hardest
fastest
best
easiest
slowest
simplest
big
coolest
hottest
coldest
hard
good
easy
fast
bigger
simple
better
slow
harder
easier
faster
slower
simpler
hot
cool
cold
hotter
cooler
colder

Motivation

- Finding hidden language relations
- Learning more from single vector representations
- Recent methods from topological data analysis (TDA)

Main goals

Deepen our understanding of word embeddings through

1. Cluster analysis
2. Prediction of word polysemy

Methods

Methods

- Analogy and cluster analysis
  - Classical word embedding model evaluation
  - Clustering and internal clustering validation algorithms
- Polysemy prediction
  - Topological polysemy
  - Geometric Anomaly Detection (GAD)
  - Intrinsic dimension estimation
  - Proposed supervised models

Analogy analysis

- Analogy test data sets
  - Syntactic, e.g. capital cities, currencies
  - Semantic, e.g. comparative, past tense
- Examples:
  - king – man ≈ queen – woman
  - Paris – France ≈ Norway – Oslo
  - good – better ≈ cold – colder
  - work – works ≈ speak – speaks

Analogy analysis

$$\text{king} - \text{man} \approx \text{queen} - \text{woman}$$

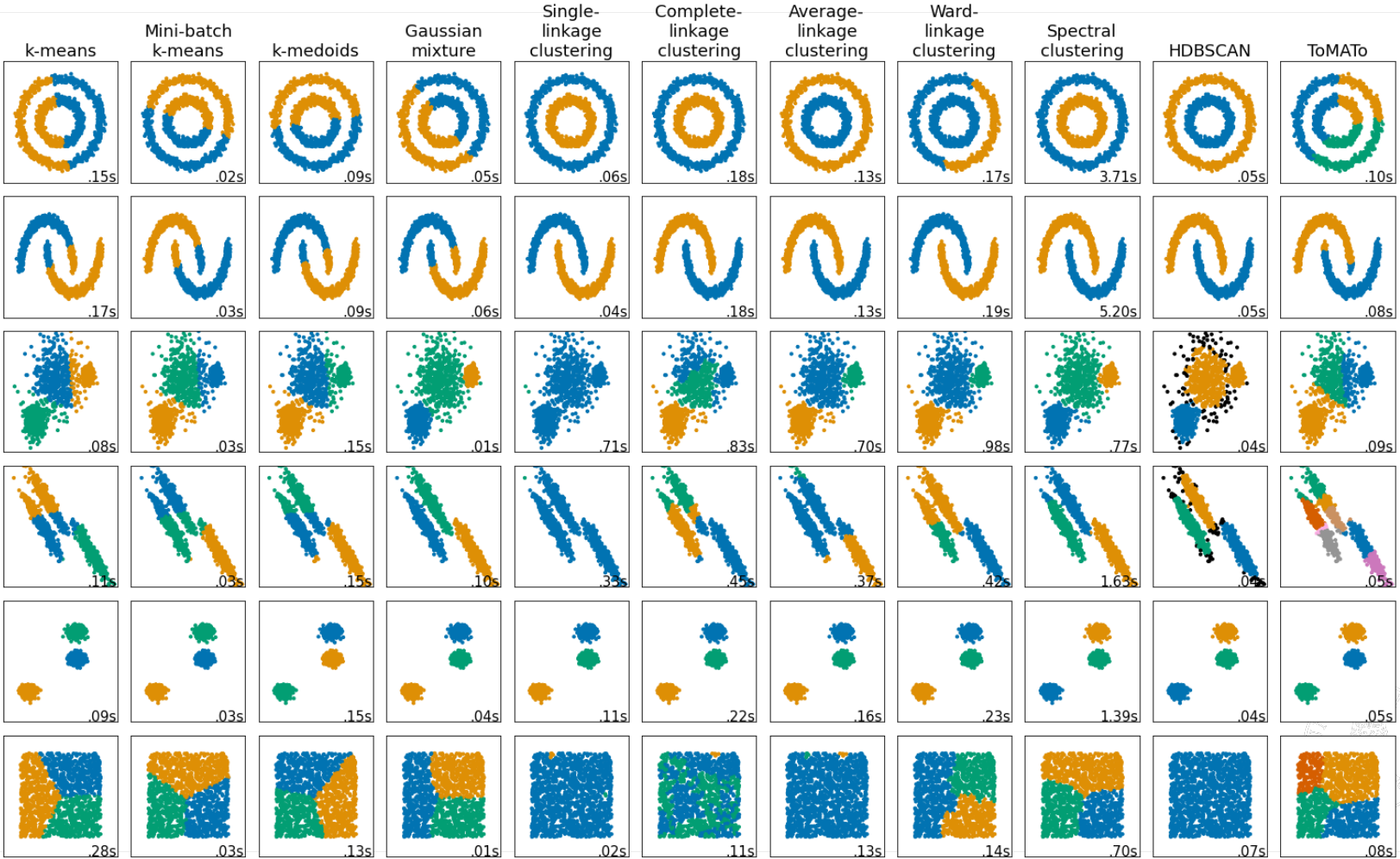
$$\text{queen} \approx \text{king} - \text{man} + \text{woman}$$

Converting to word embeddings:

$$v_{\text{queen}} \approx v_{\text{king}} - v_{\text{man}} + v_{\text{woman}}$$

Cluster analysis

- Clustering algorithms
- Internal cluster validation methods



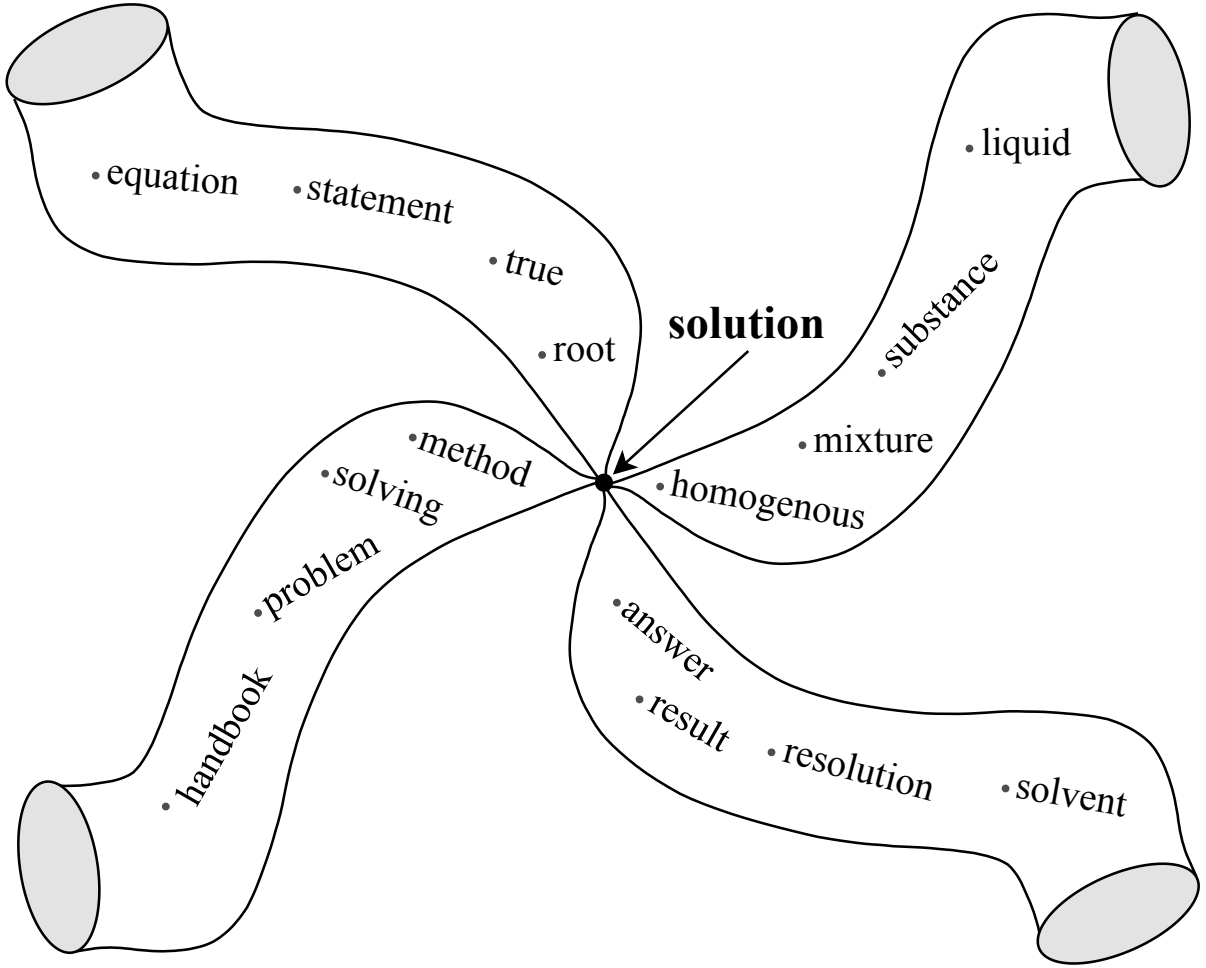
Source: https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html, downloaded and modified 9 June 2021.

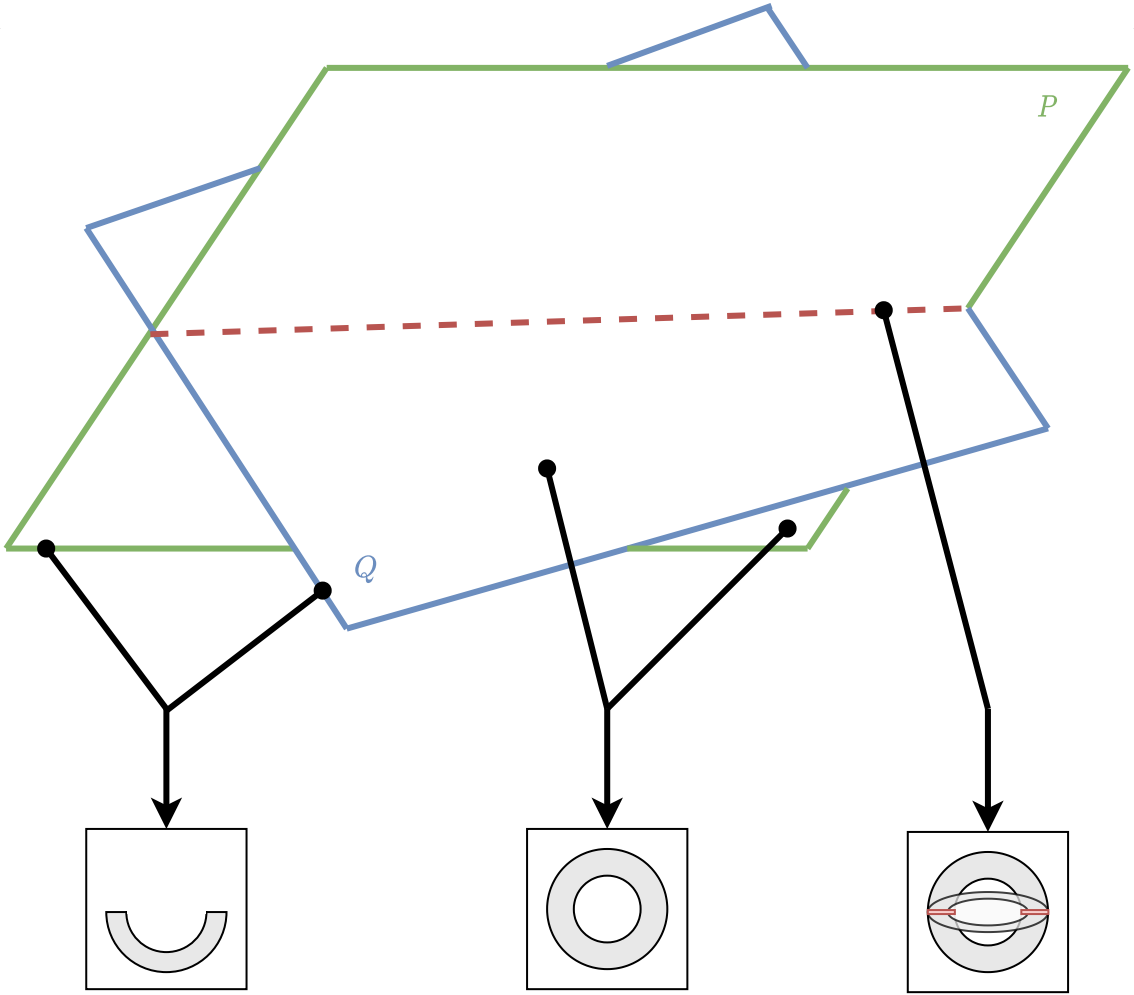
Cluster analysis

- Internal cluster validation methods
  - Silhouette Coefficient (SC)
  - Davies-Bouldin Index (DBI)
  - Caliński-Harabasz Index (CHI)

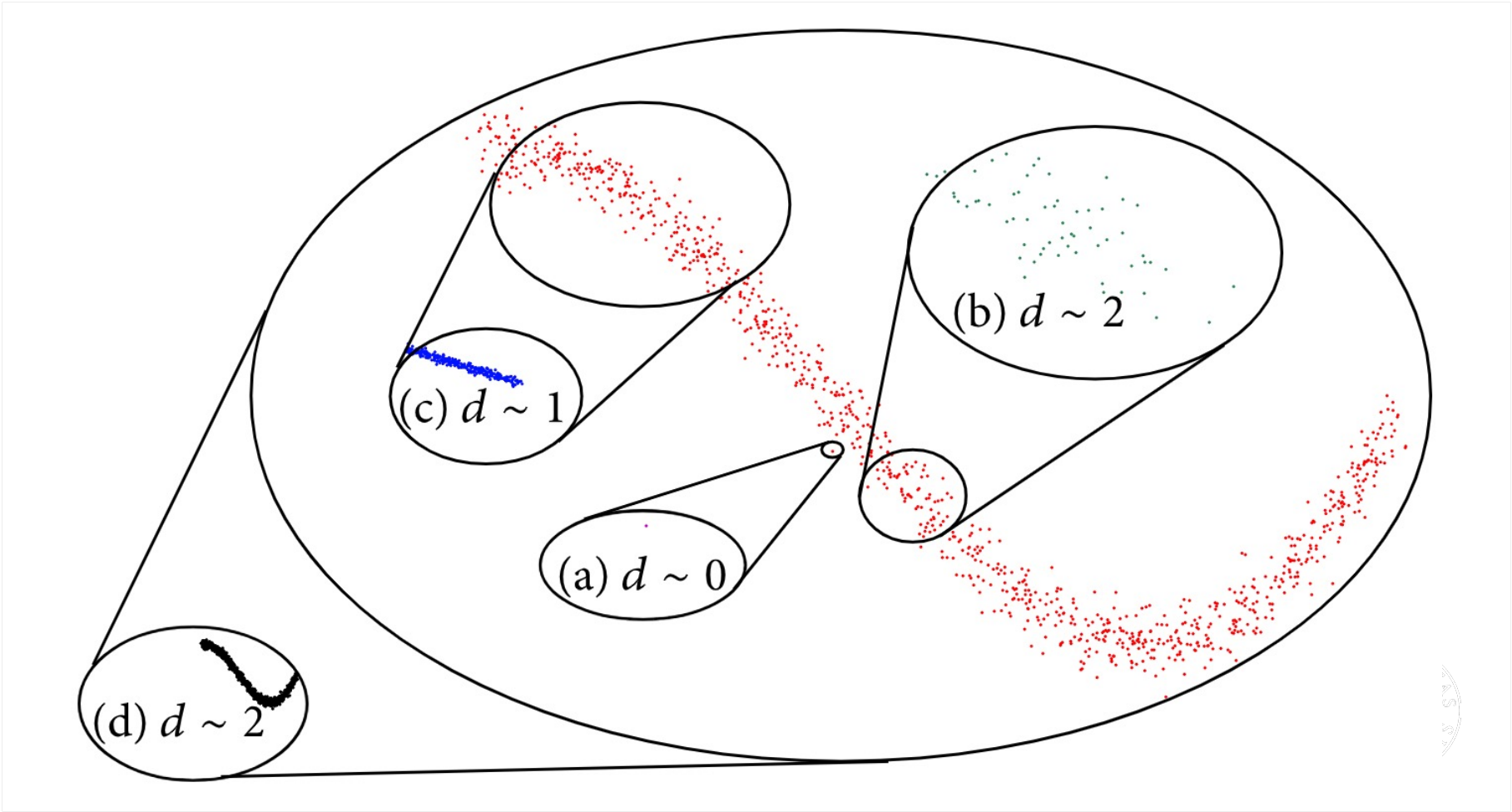
Polysemy prediction

- Topological data analysis
  - Topological polysemy
  - Geometric Anomaly Detection
- Intrinsic dimension estimation
- Regression analysis
  - Lasso regression
  - Logistic regression


equation
statement
true
root
solution
liquid
substance
mixture
homogenous
method
solving
problem
handbook
answer
result
resolution
solvent



(a) Boundary point (b) Manifold point (c) Singular point



Source: P. Campadelli et al. (2015), Figure 1: https://www.hindawi.com/journals/mpe/2015/759567.

Results

Training our word2vec models

- SGNS-enwiki (English Wikipedia)
  - ~4.4M words in vocabulary
- SGNS-semeval (SemEval-2010 task 14 data)
  - ~122k words in vocabulary
- 300 word embedding dimensionality

Analogy analysis

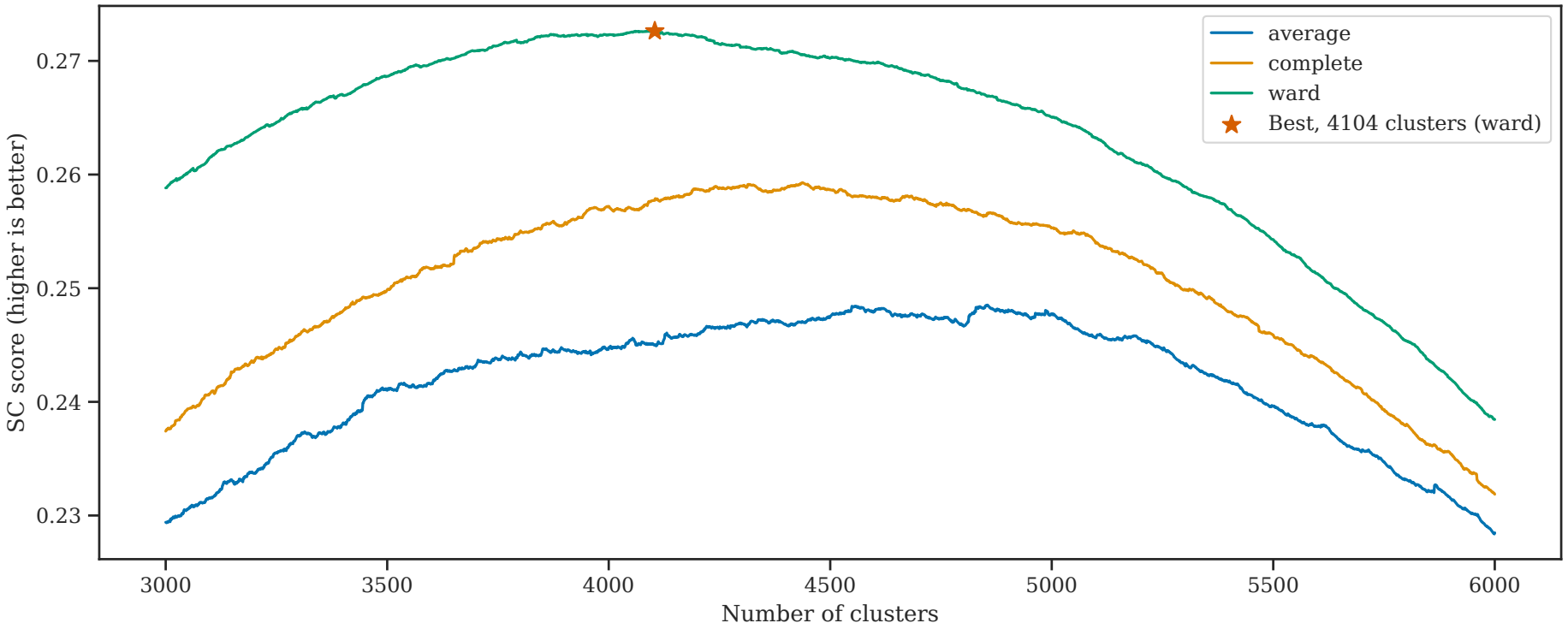
- Semantic-Syntactic Word Relationship test set (SSWR)
  - 19.5k analogy tests (semantic and syntactic)
- Microsoft Research Syntactic Analogies Dataset (MSR)
  - 8k analogy tests
- Phrase Analogy Dataset (PAD)
  - 3k phrase analogy tests

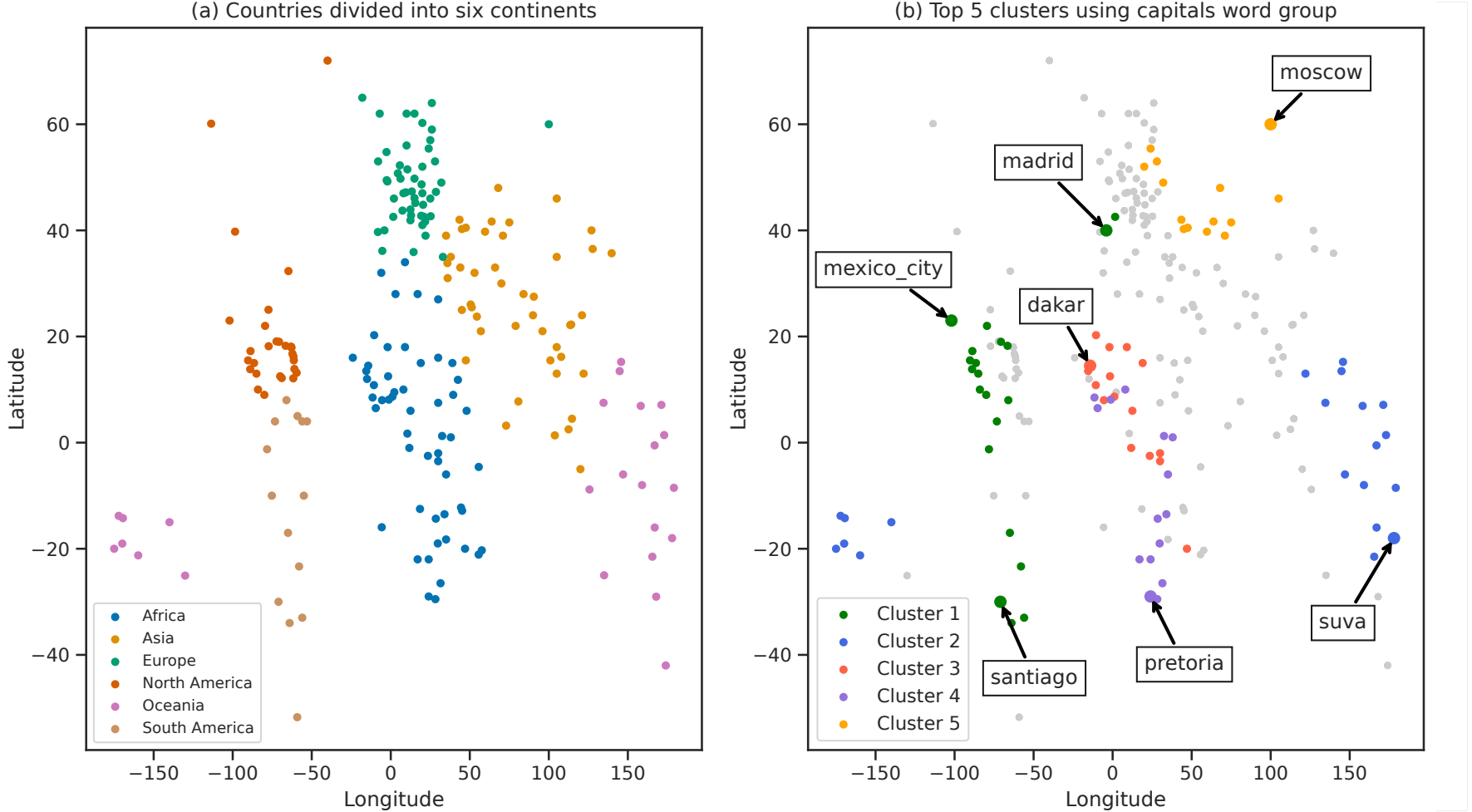
Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Model	SSWR		
	Semantic	Syntactic	Average
SG 300	55	59	57
SG 1000	66.1	65.1	65.6
NEG-15	61	61	61
RNN-1600	–	–	–
GloVe 300 42B	**81.9**	69.3	75.0
fastText	77.8	**74.9**	**76**
SGNS-enwiki	65.8	67.3	66.6

Cluster analysis

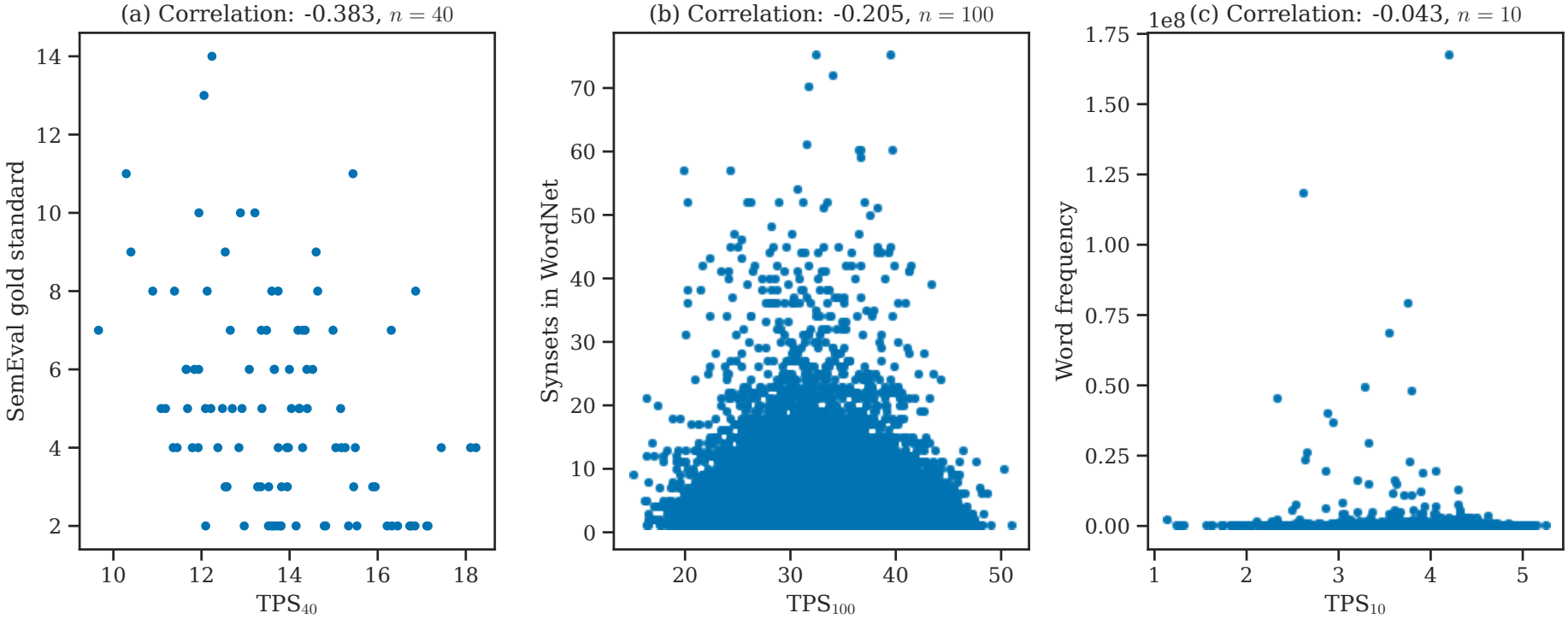
- Comparing cluster algorithms
- Internal cluster validation methods
- Analysis of
  - 4.4M → 10k most common words
  - Distinct word groups
    - Countries/capitals
    - Number words

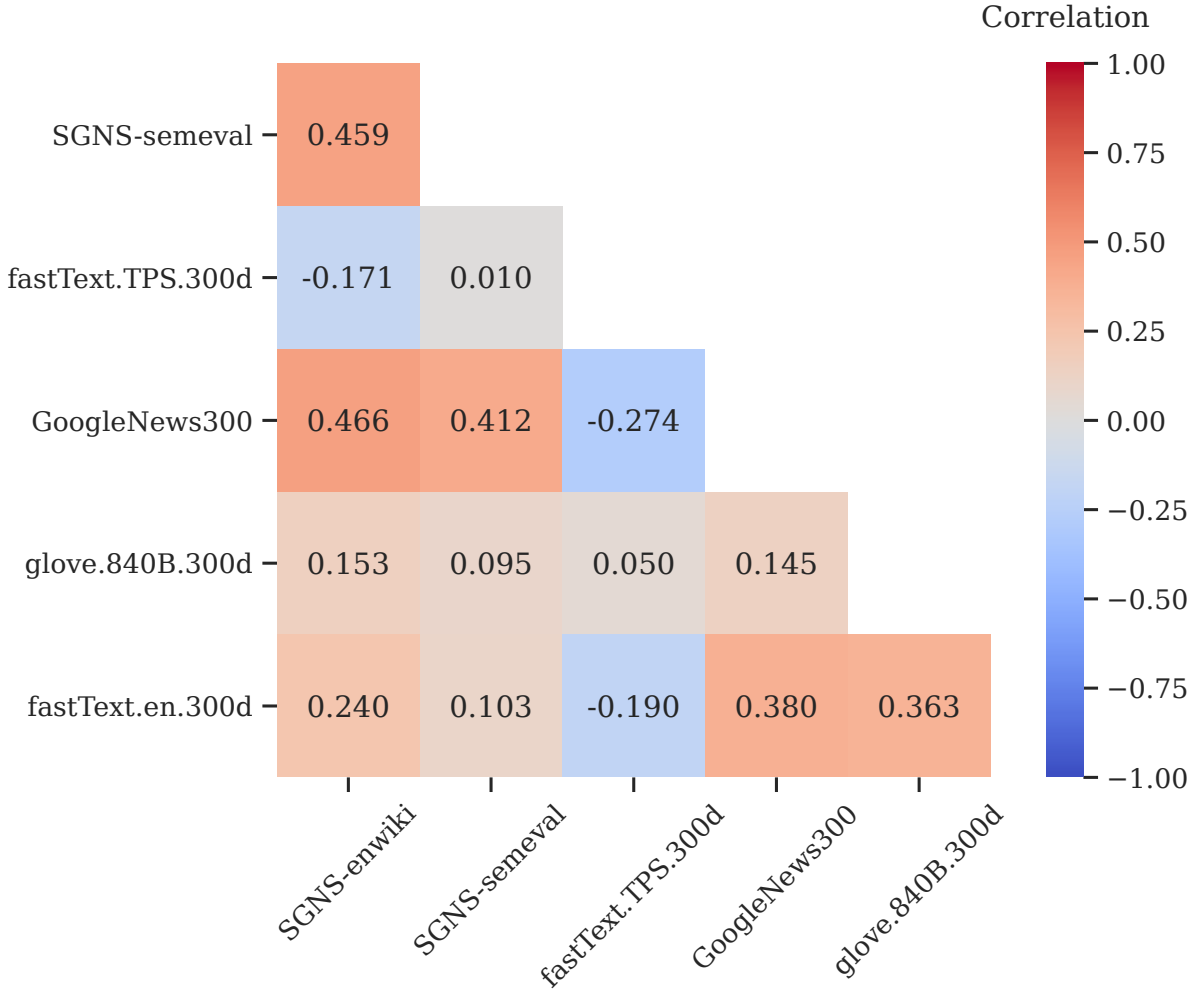

average
complete
ward
Best, 4104 clusters (ward)
SC score (higher is better)
0.27
0.26
0.25
0.24
0.23
3000
3500
4000
4500
5000
5500
6000
Number of clusters


(a) Countries divided into six continents
Latitude
Longitude
60
40
20
0
−20
−40
−150
−100
−50
0
50
100
150
Africa
Asia
Europe
North America
Oceania
South America
(b) Top 5 clusters using capitals word group
moscow
madrid
mexico_city
dakar
santiago
pretoria
suva
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5

Topological polysemy

- Word embedding models
- SemEval-2010 task 14
  - 100 polysemous words
  - Gold standard (GS)
- Correlation with
  - Gold standard
  - Number of WordNet synsets
  - Word frequency


(a) Correlation: -0.383, $n = 40$
SemEval gold standard
TPS_{40}
(b) Correlation: -0.205, $n = 100$
Synsets in WordNet
TPS_{100}
(c) Correlation: -0.043, $n = 10$
1e8
Word frequency
TPS_{10}


Correlation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SGNS-semeval
fastText.TPS.300d
GoogleNews300
glove.840B.300d
fastText.en.300d
0.459
-0.171
0.010
0.466
0.412
-0.274
0.153
0.095
0.050
0.145
0.240
0.103
-0.190
0.380
0.363
SGNS-enwiki
SGNS-semeval
fastText.TPS.300d
GoogleNews300
glove.840B.300d

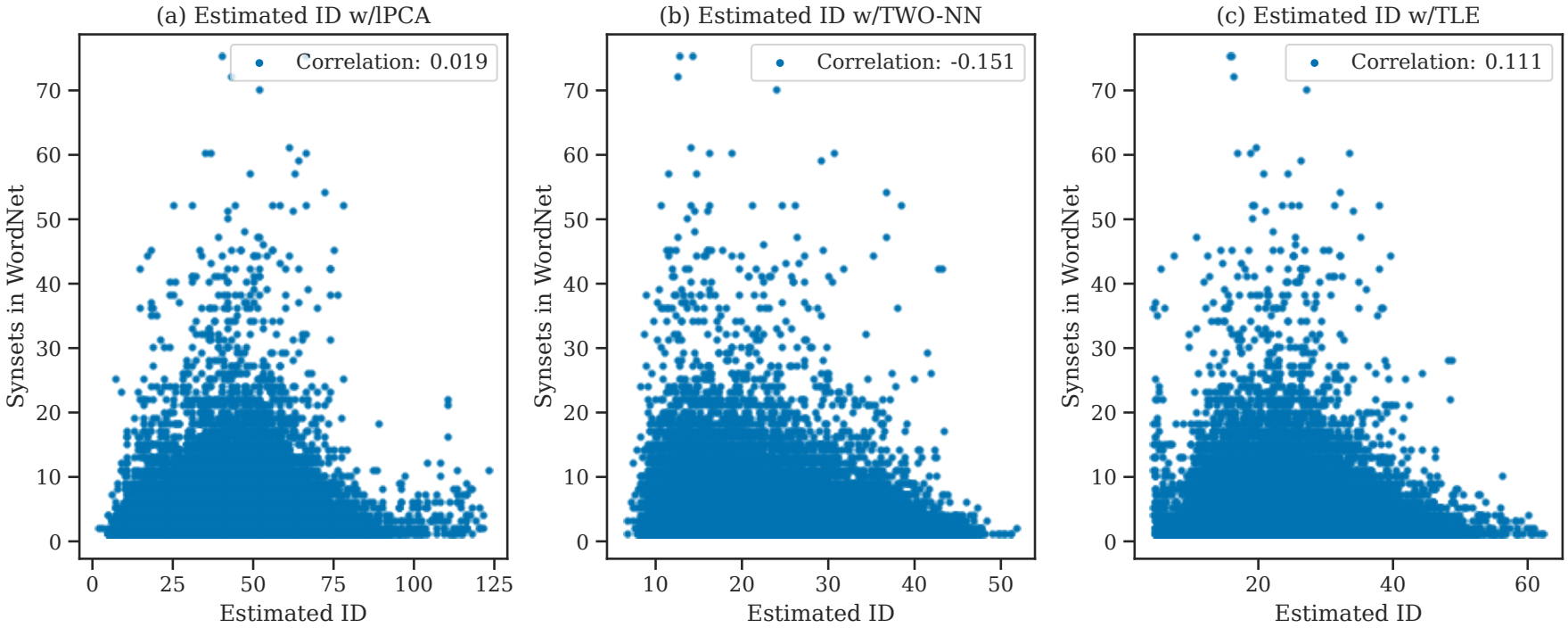
Geometric Anomaly Detection (GAD)

- WordNet SGNS-enwiki word embeddings
- Singular words reflected by singular GAD group

	GAD group			
	Manifold	Boundary	Singular	*Sum*
Number of monosemous words	4640	86731	4161	95532
Number of polysemous words	634	47902	344	48880
Sum	5274	134633	4505	144412

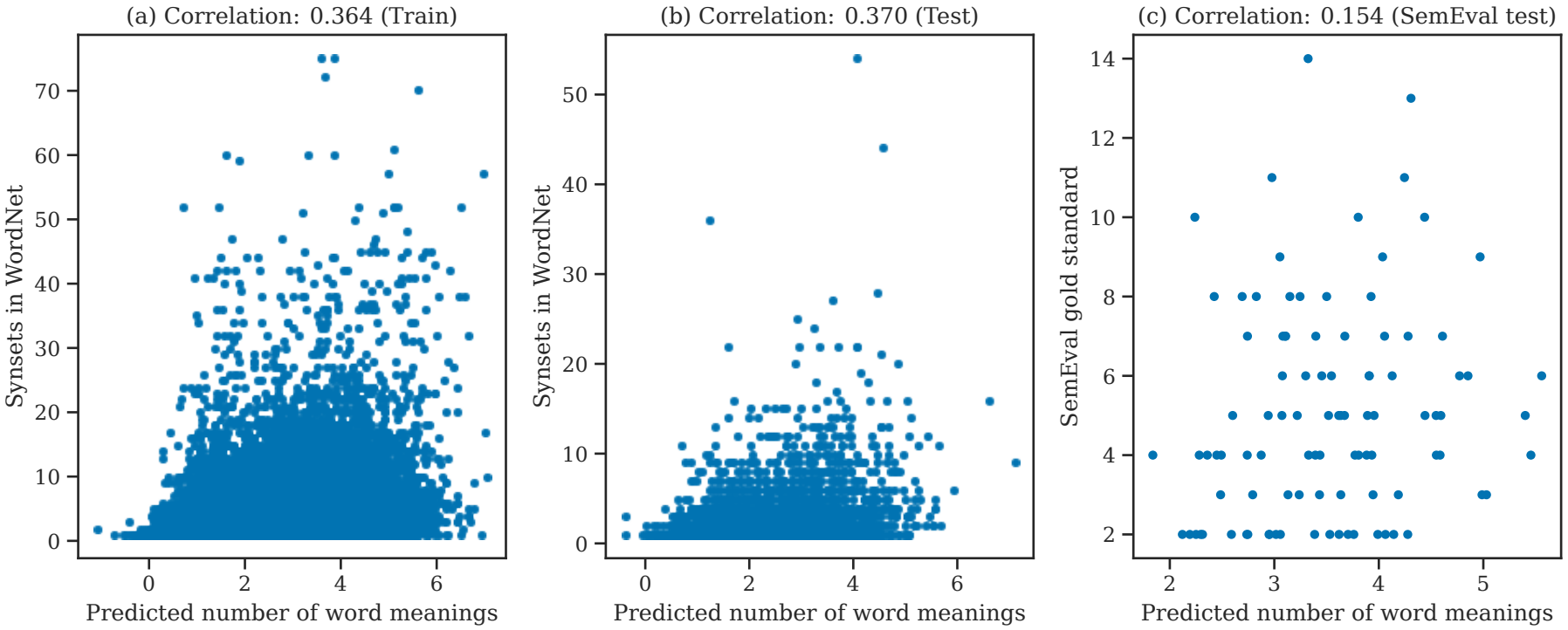
Intrinsic dimension (ID) estimation

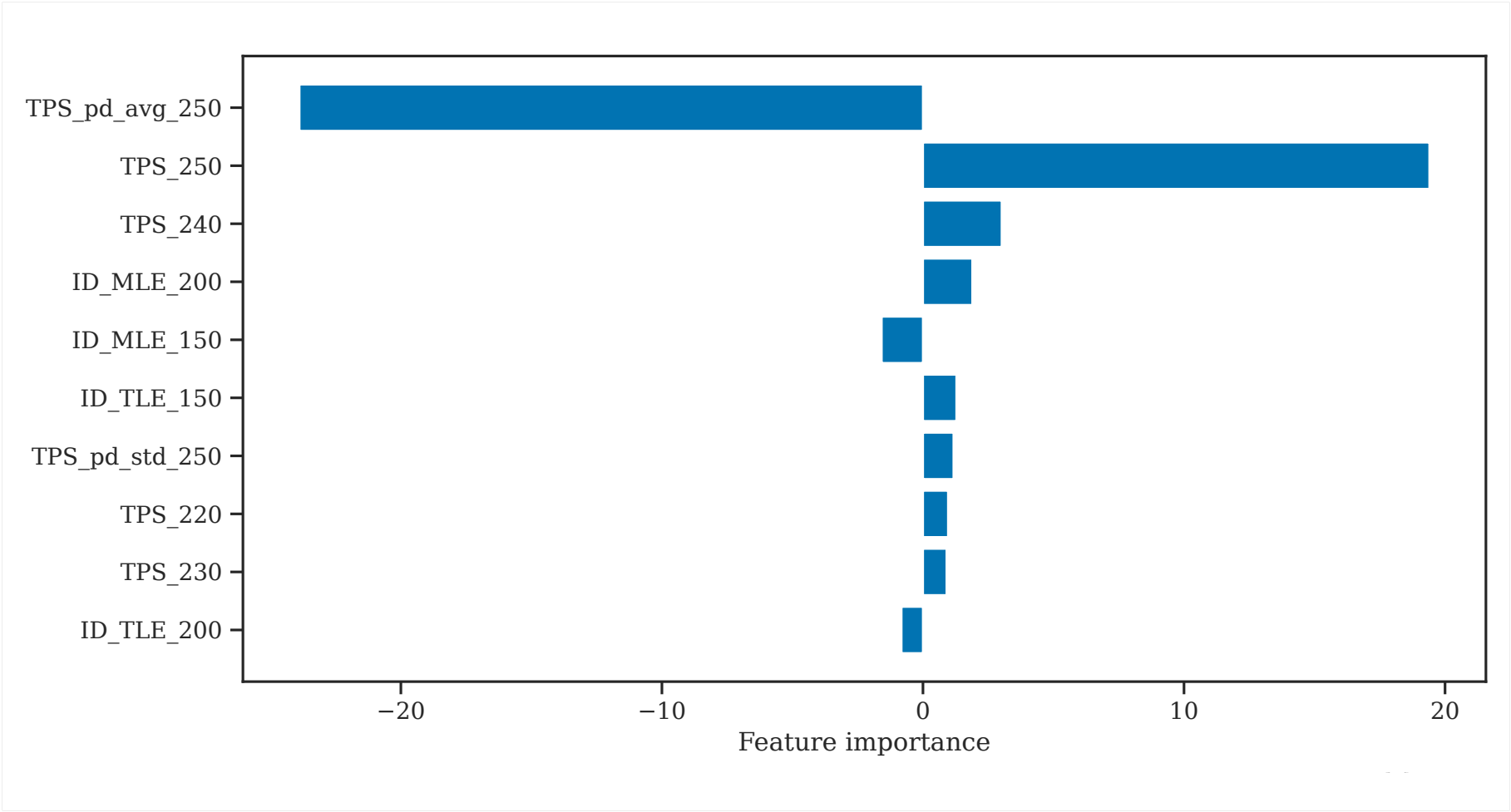
- Five ID estimation methods
- Relation to number of word meanings

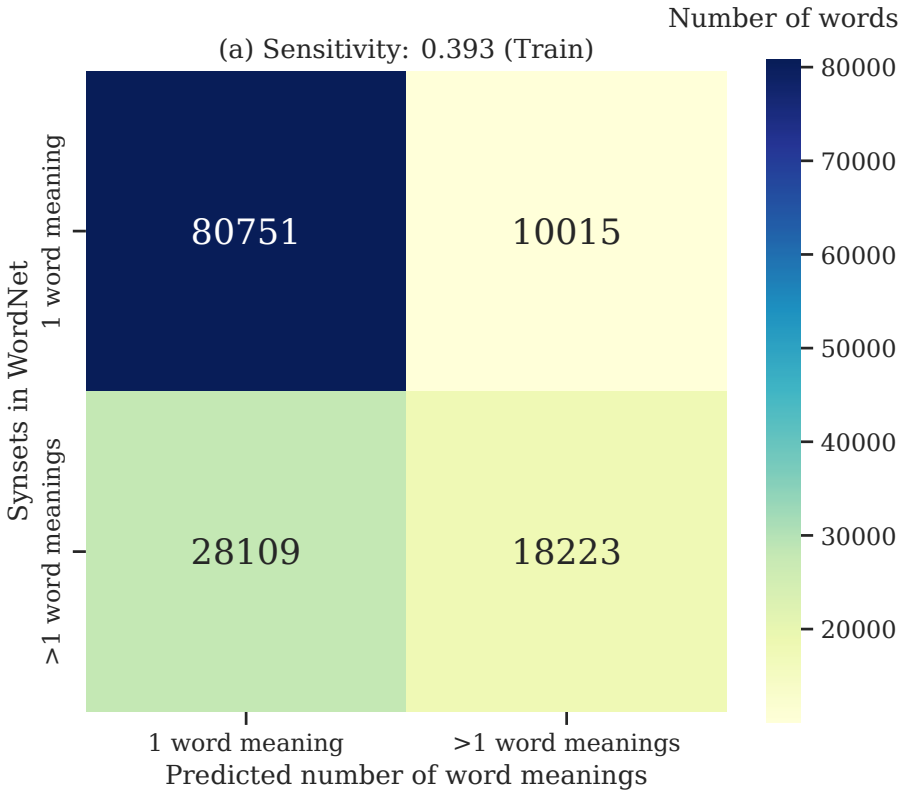

(a) Estimated ID w/lPCA
Correlation: 0.019
Synsets in WordNet
Estimated ID
(b) Estimated ID w/TWO-NN
Correlation: -0.151
Synsets in WordNet
Estimated ID
(c) Estimated ID w/TLE
Correlation: 0.111
Synsets in WordNet
Estimated ID

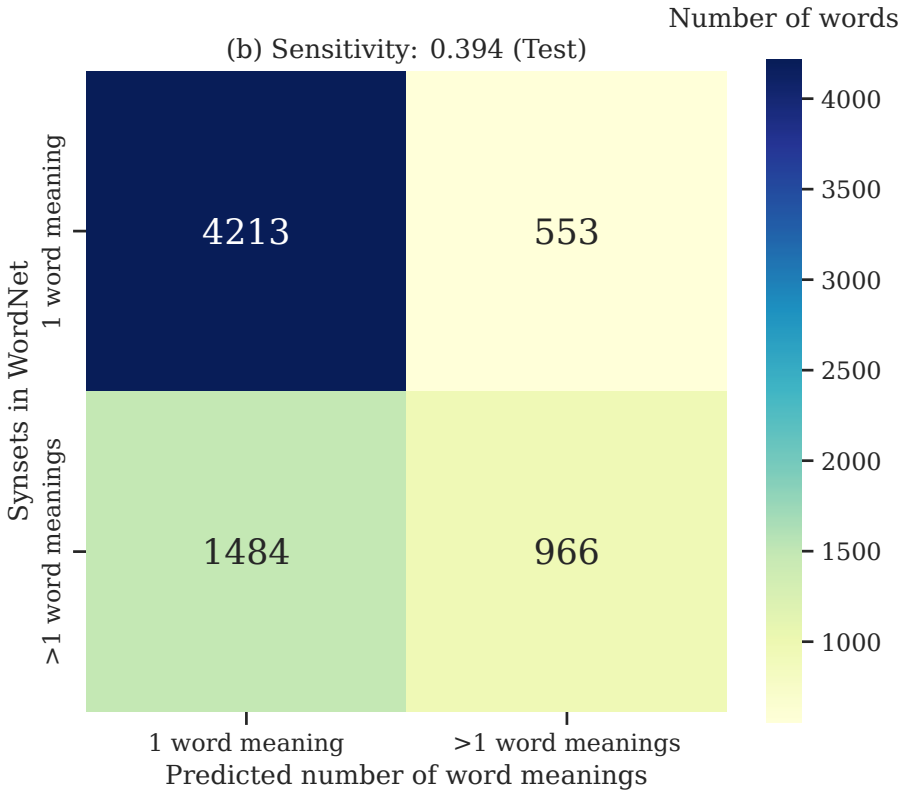
Supervised polysemy prediction

- WME- and BWME-enwiki models
  - Lasso and logistic regression
  - WordNet SGNS-enwiki word embeddings
- Features from TPS, GAD and ID estimation
- 95% train / 5% test split
- 20-fold cross validation on ℓ_1-hyperparameter


(a) Correlation: 0.364 (Train)
Synsets in WordNet
Predicted number of word meanings
(b) Correlation: 0.370 (Test)
Synsets in WordNet
Predicted number of word meanings
(c) Correlation: 0.154 (SemEval test)
SemEval gold standard
Predicted number of word meanings


TPS_pd_avg_250
TPS_250
TPS_240
ID_MLE_200
ID_MLE_150
ID_TLE_150
TPS_pd_std_250
TPS_220
TPS_230
ID_TLE_200
−20
−10
0
10
20
Feature importance


(a) Sensitivity: 0.393 (Train)
Number of words
80000
70000
60000
50000
40000
30000
20000
Synsets in WordNet
1 word meaning
>1 word meanings
80751
10015
28109
18223
1 word meaning
>1 word meanings
Predicted number of word meanings


(b) Sensitivity: 0.394 (Test)
Number of words
4000
3500
3000
2500
2000
1500
1000
Synsets in WordNet
1 word meaning
>1 word meanings
4213
553
1484
966
1 word meaning
>1 word meanings
Predicted number of word meanings

Conclusion and Future Work

Conclusion

- Clustering of word embeddings deepen our understanding
- Topological polysemy yield inconsistent results

Future work

- Looking at other word embedding models
- Vectorization of GAD persistence diagrams
- GAD manifold dimension parametrization
- Word sense disambiguation problem