Automated, Corpus- and Usage-Based Semantic Classification of Word Class using Word Embeddings

Martin Schweinberger and Chang-Hao (Howard) Luo

The University of Queensland

Starting Point

Meaning obviously represents fundamental property of language and semantics impact all levels of linguistic analysis.

Including semantics in linguistic analyses is, however, hampered by the fact that the annotation of semantics is laborious, time-intensive, and often subjective.

This study aims to explore options that use a usage-based, automated method for annotating semantics so that meaning aspects are easier to integrate into linguistic analyses.

Overview of the study we present today

Aim

- Develop a semantic classification system of adjectives using word embeddings

Motivation

- Studies in LVC (e.g., Tagliamonte, 2007) suggest that the semantics of adjectives impact amplifier choice in ongoing language change scenarios

Outline

- Background
  - Existing feature-based classifications
  - What are word embeddings, UMAP, and K-means clustering
- Methodology: what we have done
- Results and evaluation
- Issues, Limitations and where from here?

Background

Background

Existing feature-based semantic classification systems

- **Typology-based classification** (Dixon 1977)
  Aims to determine sematic features that underlie a universal, language-independent classification of semantic groups of adjectives based on syntactic and morphological properties of adjectives
- **Corpus/Distribution-based classification** (Biber et al. 2007)
  Uses the Longman Spoken and Written English Corpus to extract frequencies of structure patterns in order to define grammatical properties of words, including using semantic features to group adjectives.
- **Automated, computational classification** (USAS, distribution- and context-based) (Rayson, Archer & Piao, 2004)
  Uses the UCREL semantic analysis system (USAS) to automatically assign semantic tags based on a combination of part-of-speech (POS) tagging, a lemmatiser, and semantic tagging (Rayson, Archer & Piao, 2004)

Typology-based classification (Dixon 1977)

Semantic classes

- Dimension (size, length/width): *big, little, long, wide, thin*
- Physical Property: denoting physical attributes: *hard, light, smooth, sweet*
- Colour: *red, blue, black, white*
- Human Propensity (emotions or personality traits): *jealous, kind, dumb, happy, generous*
- Age: *new, old, young*
- Value (denoting judgement): *good, bad, proper, excellent, poor*
- Speed: *fast, quick, slow*

Corpus/Distribution-based classification (Biber et al. 2007)

Descriptors (denote physical features or characteristics)

- Colour (denoting colour or brightness): *black, white, dark, light*
- Size/weight (denoting size or weight): *big, deep, heavy, tall*
- Time (denoting frequency or age): *old, annual, late, new*
- Evaluative/emotive (denoting judgements and emotions): *good, worse, lovely, poor*
- Miscellaneous descriptives (denoting physical or other properties): *cold, complex, hard, private, strong*

Classifiers (categorise in relation to modified noun)

- Relational/classificational (delimits reference of nouns): *additional, final, left, original, public*
- Affiliative (national or religious reference): *American, Asian, Christian, Muslim*
- Topical (relation to a subject of area): *commercial, industrial, mental, political*

USAS (Rayson, Archer & Piao, 2004)

UCREL semantic analysis system

Assigns words and multiword expressions to one of 21 semantic fields.

Steps of the sematic tagging:

- POS Tagging: Choose the most commonly associated part of speech tag.
- Likelihood Ranking: Prioritize the most frequently associated meaning.
- Domain of Discourse: Adjust ranking based on the topic context (e.g., 'battered' in food vs. violence context).
- Text-Based Disambiguation: Maintain consistent meaning throughout the text.
- Contextual Rules: Use surrounding context to provide meaning (e.g., 'savings account' vs. 'incident account').
- Local Probabilistic Disambiguation: Use surrounding evidence to determine the correct tag (e.g., 'financial' context for words like 'bank', 'overdrawn', 'money').

A	General & Abstract Terms
B	The Body & the Individual
C	Arts & Crafts
E	Emotional Actions, States & Processes
F	Food & Farming
G	Government & the Public Domain
H	Architecture, Building, Houses & the Home
I	Money & Commerce
K	Entertainment, Sports & Games
L	Life & Living Things
M	Movement, Location, Travel & Transport
N	Numbers & Measurement
O	Substances, Materials, Objects & Equipment
P	Education
Q	Linguistic Actions, States & Processes
S	Social Actions, States & Processes
T	Time
W	The World & Our Environment
X	Psychological Actions, States & Processes
Y	Science & Technology
Z	Names & Grammatical Words

Issues of existing semantic classification systems

- **Time consuming**
  Manual annotation of semantic classes are very time-intensive
- **Subjective**
  Manual annotation can have low inter-rater reliability and show inconsistencies hindering reproducibility
- **Non-intuitive semantic categories/classes**
  Automated annotation has high access/accuracy, is easy to implement, and produces replicable results but the semantic classes appear un-intuitive and not aligned with existing sematic classes

Our approach

- We try to build on the manual semantic classification systems and use word embeddings as the basis of an alternative automated classification system that can be used by the research community

What are Word Embeddings?

Usage-based measure of semantic similarity (Chandrasekraran & Mago, 2021; Haripse et al., 2022)

> Word embeddings are a way to represent words as numerical vectors so that words with similar meanings are closer together in this numeric space. They are used in NLP to give computers a way to understand and process human language by capturing the relationships between words

Words with similar meaning have similar numeric vectors (because they occur in similar contexts) and would thus be displayed close to each other in two-dimensional space.

What are Word Embeddings?

Word embeddings are long vectors of numbers representing word types (or tokens).
Example:

cat	(5,	7,	1,	0,	8)
dog	(5,	8,	1,	0,	7)
table	(1,	0,	8,	9,	2)
chair	(1,	1,	7,	8,	2)

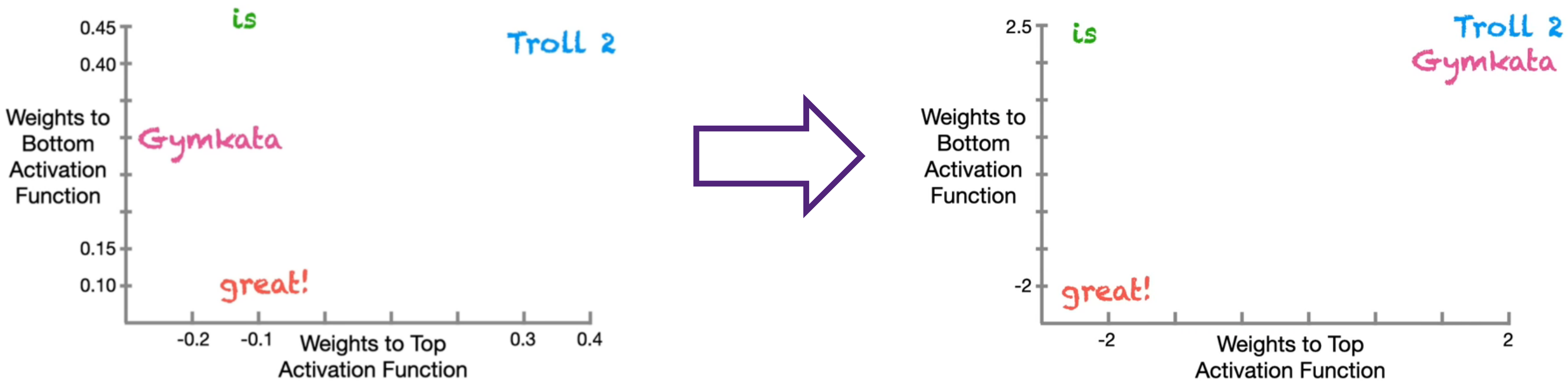
Example

(1) *Troll2 is great!*

(2) *Gymkata is great!*

Initially words are assigned a vector of random numbers. A random set of words is chosen: if words have the same context (surrounding words), their numbers are made more similar (over many iterations)

Source: StatQuest,2023

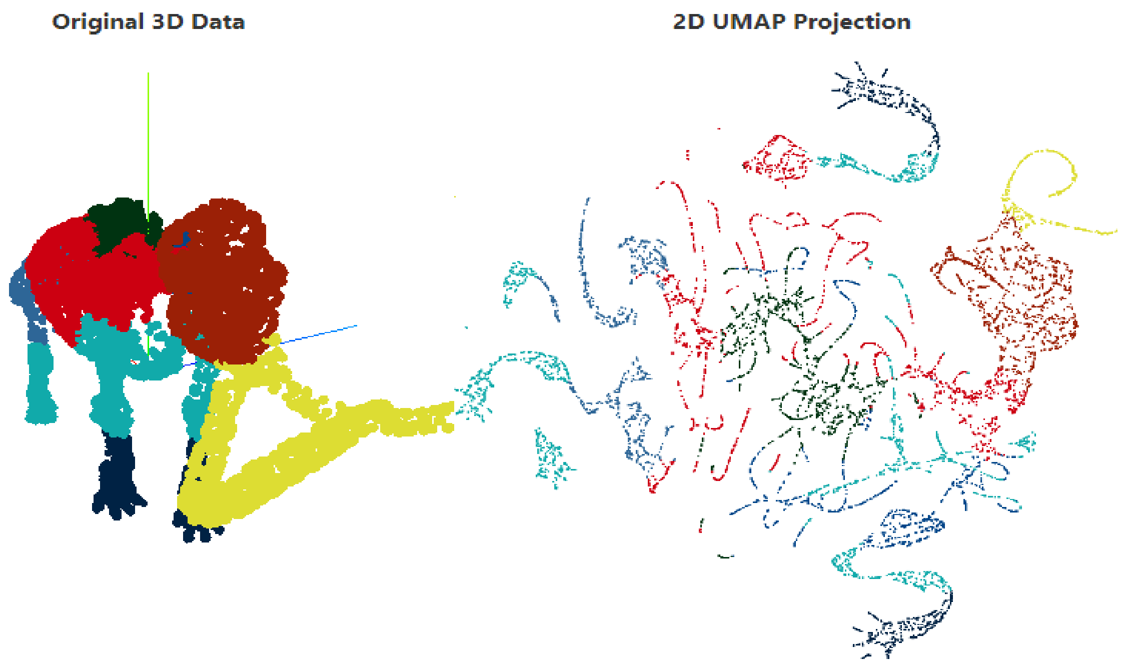


What is UMAP?

Uniform Manifold Approximation and Projection (UMAP) (see McInnes, Healy and Melville, 2020) is a statistical procedure used to reduce the number of dimensions from high-dimensional to low-dimensional data with minimal information loss.

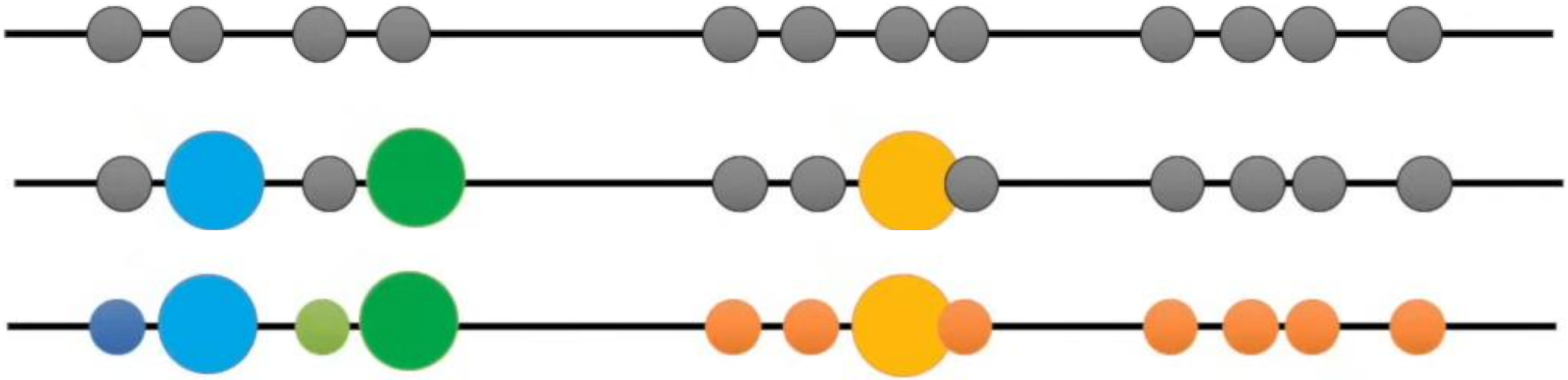
The reduction of dimensionality renders complex data to be more readily analysable and visualizable.

It preserves both local and global structure of the data (groupings, distance, and patterns), helping to see relationships in a simpler, two-dimensional or three-dimensional space.



What is K-means Clustering?

Agglomerative clustering method (Franti, Virmajoki, & Hautamaki, 2006) used to identify what clusters elements belong to. The number of clusters (K) is set by the researcher. All elements are then assigned to the cluster they are closest to (closeness depends on what method is used).



Source: StatQuest,2023

Data and Methodology

Methodology: what have we done?

Data

Pre-existing word embeddings (trained on news texts): Gigaword 5th edition corpus (Parker et al., 2011)

- vocabulary size of 292,479 types (each with a vector of 300 dimensions)
- generated using the Gensim Continuous Skip-gram algorithm

Processing

POS-tagging with UDPipe (Wiffels, 2021)

- Extracted 4044 adjective types

Dimension reduction with UMAP (to reduce 300 to 2 dimensions)

- UMAP: better at retaining both local and global structure during dimensionality reduction compared to similar methods (t-SNE)
- Two methods:
  - perform UMAP after clustering (UA)
  - perform UMAP before clustering (UB)

Methodology: what have we done?

KNN Clustering

- Checked different numbers of clusters (K): here we present the results for 8 and 12 clusters
- Number of clusters: automated evaluations did not provide meaningful guidance
- Number of clusters should be comparable to existing classifications (Dixon (1977) and Biber et al. (2007))
- 8 clusters did not produce a fine-grained enough solution

Evaluation

- How did our results hold up against existing classifications (Biber et al. 2007)?
- Eye-balling classification (manual checks)
- Coherence metrics: How consistent were our results?
  - draw sample from each cluster to determine category type
  - generate new set of words based on category and check cluster allocation
  - generate Confusion Matrix and calculate coherence by comparing actual and predicted clustering

Methodology: what have we done?

Evaluation

- Coherence metrics: How consistent were our results?

Cluster X	Feature
Red	Color
Blue	Color
Purple	Color
Salty	Misc.
Green	Color

→ Color →

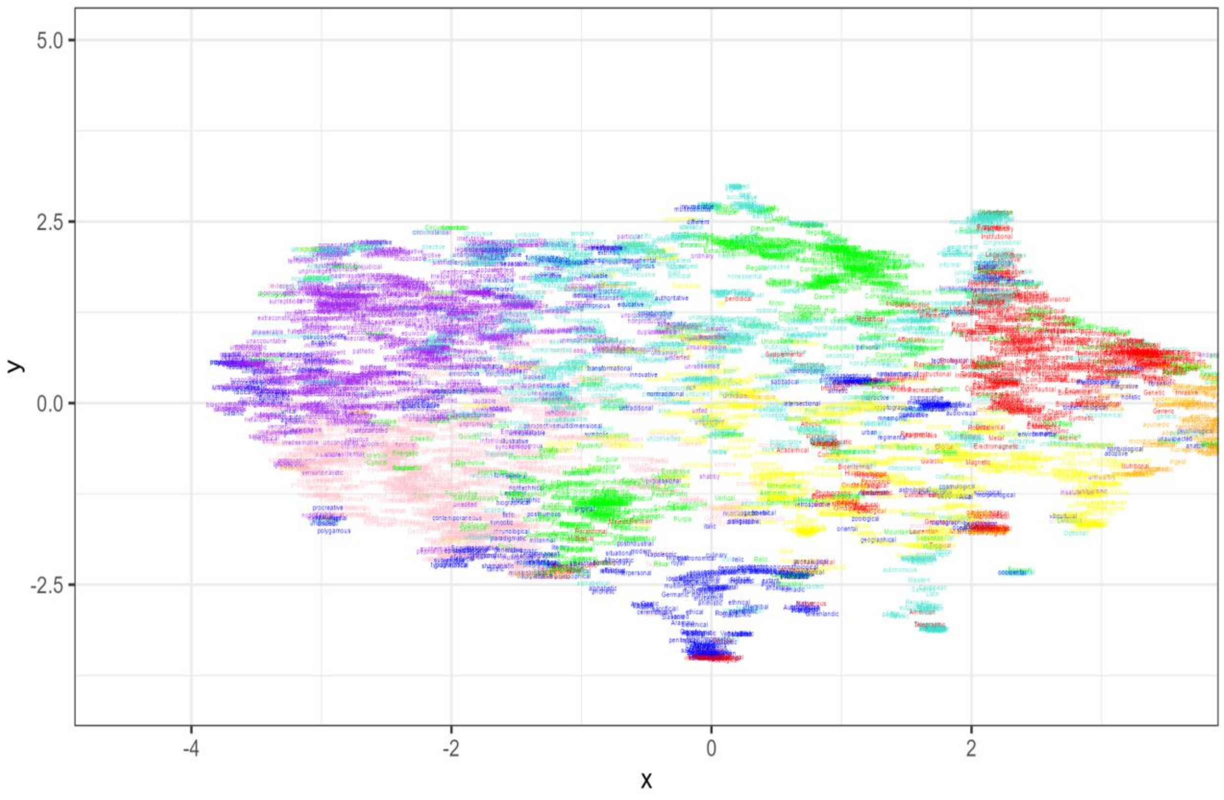
Comparison	Cluster
Yellow	X
Aqua	X
Maroon	B
Teal	X
Violet	B

Accuracy
3/5 = 60 %

Results and Evaluation

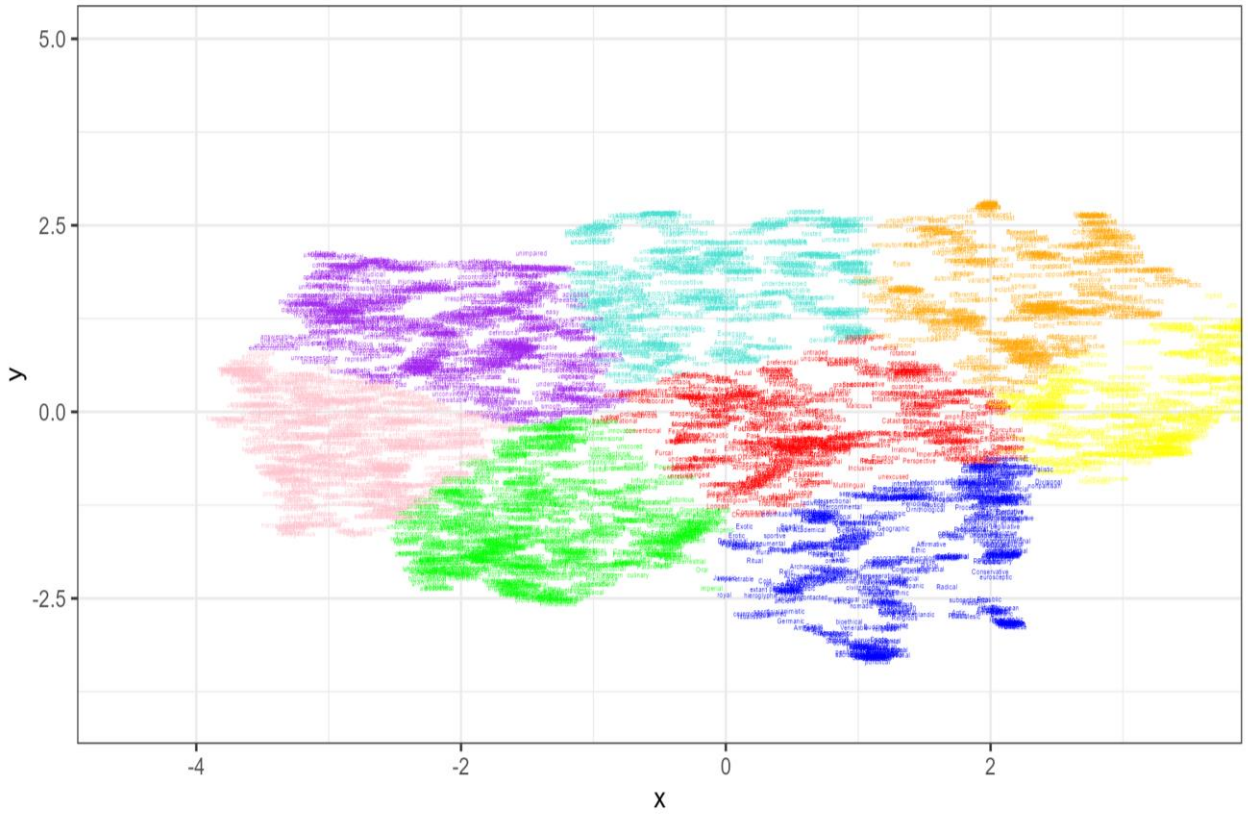
Clustering before and after Dimension Reduction

Results: UMAP before KNN (8K)



Average Consistency = 63%

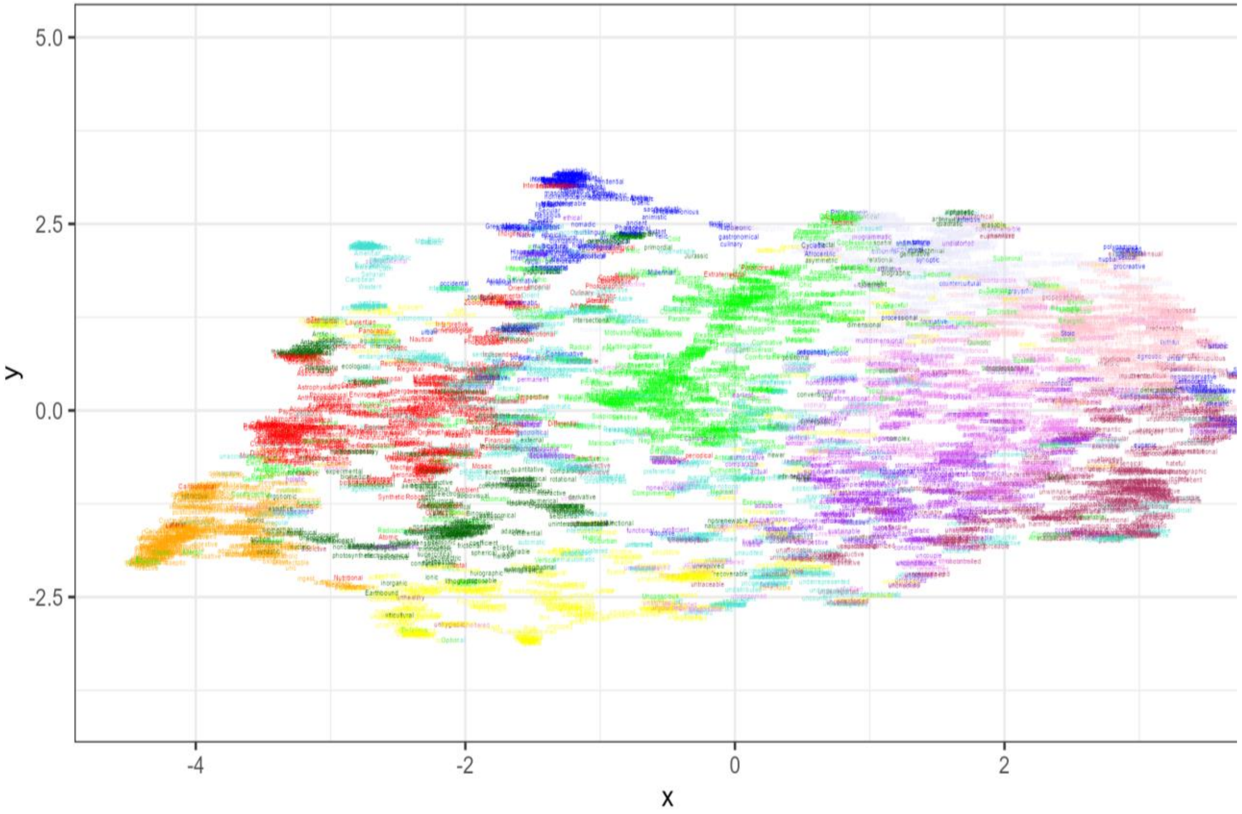
Results: UMAP after KNN (8K)



Average Consistency = 67%

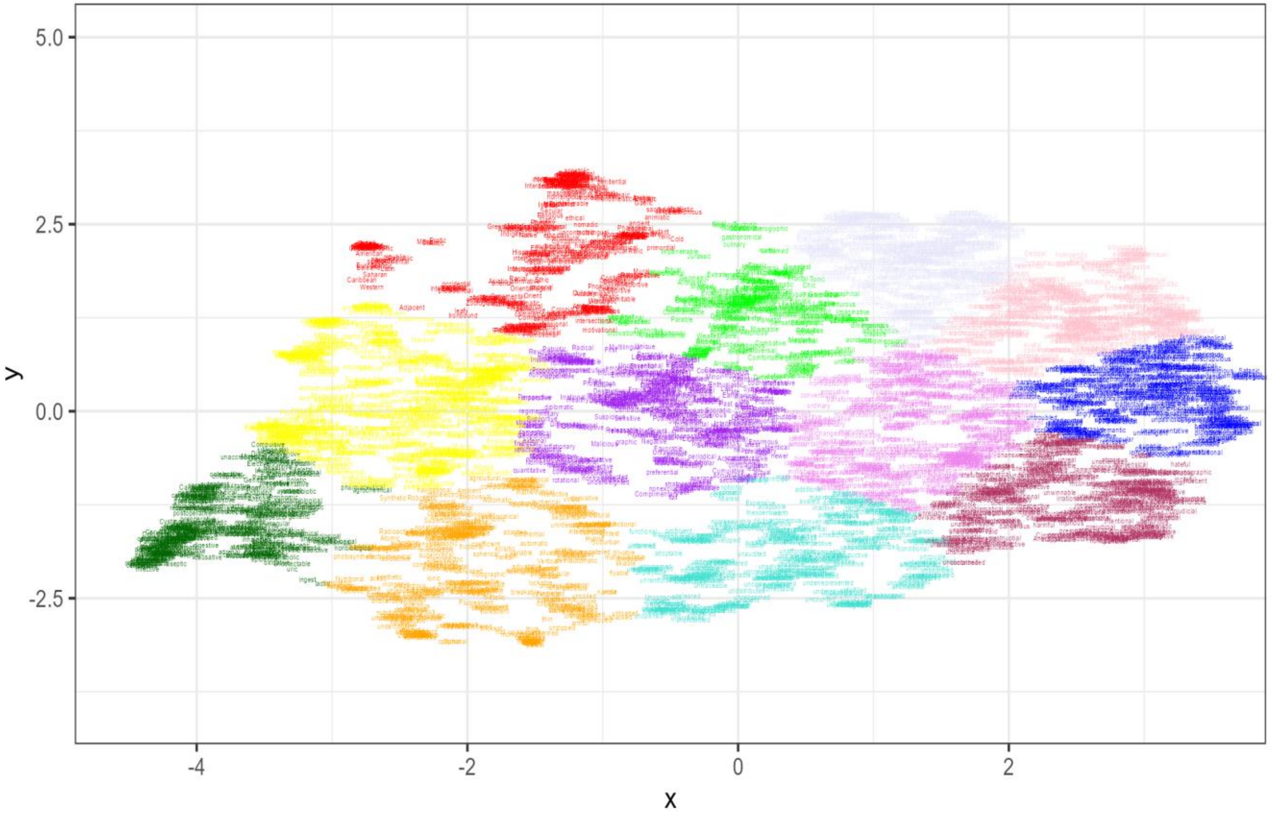
Clustering before and after Dimension Reduction

Results: UMAP before KNN (12K)



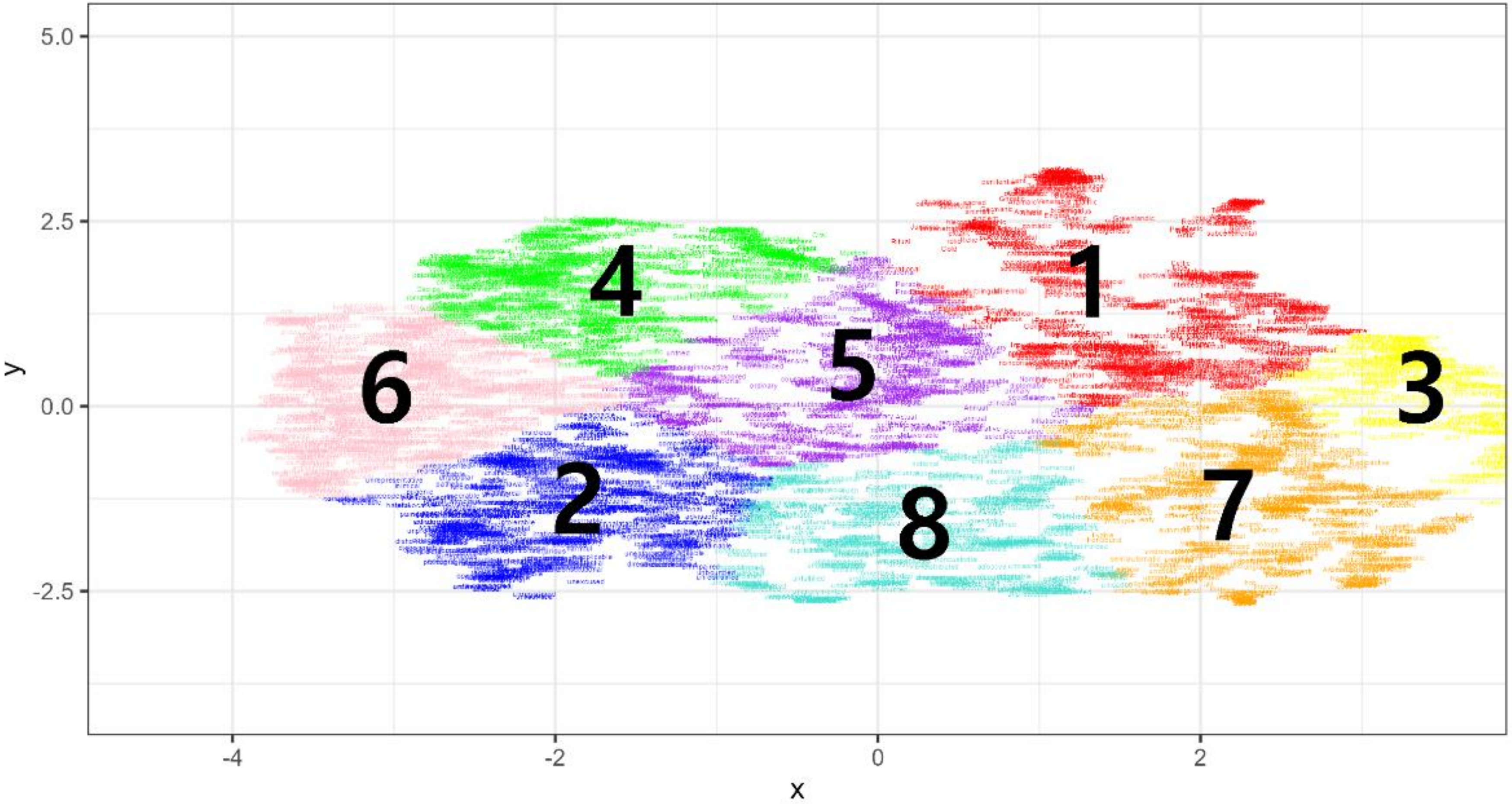
Average Consistency = 50%

Results: UMAP after KNN (12K)



Average Consistency = 52%

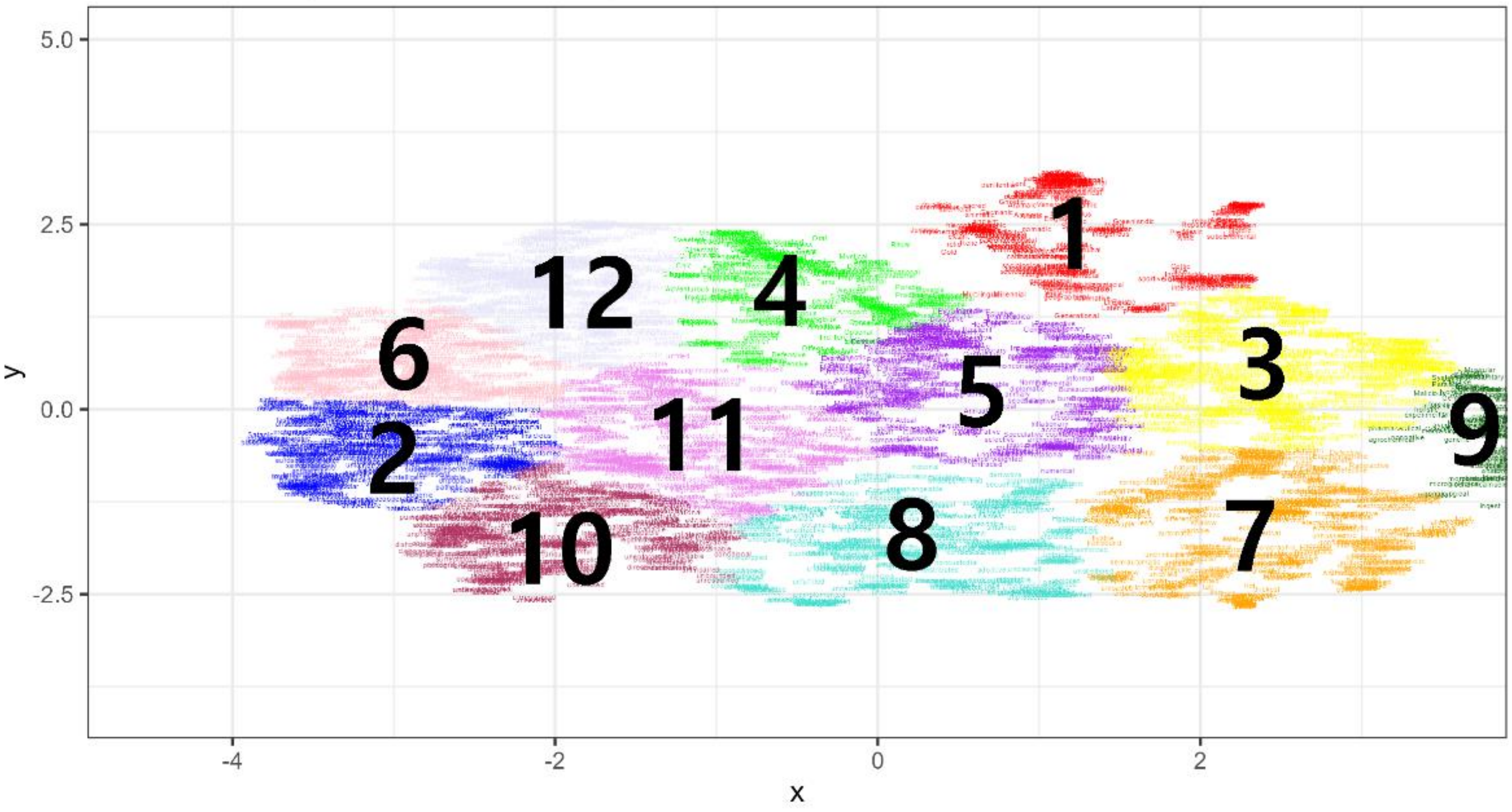
Clustering Solution (K8)



Average Consistency = 67%

Id	Label	Example
1	Affiliative (nation/state/region)	American, Japanese, Islamic
2	Evaluative (judgemental)	hard, bad, possible, wrong
3	Medical	mental, genetic, spinal
4	Art/Genre related	popular, beautiful, romantic
5	Descriptive (misc)	expensive, cheap, electronic
6	Human traits	loyal, glad, faithful
7	Domestic news-related	several, high, military
8	Descriptives (non-human)	big, small, strong, largest

Clustering Solution (K12)



Average Consistency = 52%

Id	Label	Example
1	Affiliative (nation/ region)	American, Japanese, Islamic
2	Neg. human trait	concerned, violent, angry
3	Domains	industrial, public, academic
4	Descriptive (biographic)	former, influential, youngest,
5	Descriptive (news-related)	several, high, military
6	Pos. human traits	funny, romantic, charismatic
7	Descriptives (phy./ product)	expensive, cheap, electronic
8	Descriptives (non-human)	close, dead, worth
9	Medical	medical, mental, genetic
10	Evaluative (sit.)	possible, difficult, criminal
11	Evaluative(pos.)	good, best, strong
12	Art/Genre-related	romantic, classical, epic

Evaluation of Semantic Categories

Some were easily recognisable and had high accuracy, but other categories were difficult to categorise and lacked consistency.

Selected findings

- **Easy and accurate categorisation**
  Medical (83% acc.): *medical*, *physical*, *mental*, *fatal*, *clinical*, *spinal*, *surgical*
- **Mixing** (categories combining classes described in Biber et al., 2007)
  Relational and size (79% acc.): *last*, *many*, *next* , *second*, *least*, *big*, *small*, *high*, *large*
- **Over-generalisation**
  Negative Evaluative (50% acc.): *clear*, *bad*, *alleged*, *wrong*, *criminal*, *dangerous*, *worse*, *impossible*, *unclear*
- **Uncategorisable**
  Topical, Relational, Miscellaneous: *Last*, *First*, *High*, *Special*, *Free*, *Cold*, *Best*, *Official*, *Heavy*, *Domestic*

Category	Accuracy (%)
Affiliative + Topical	33.3
Descriptive (phy.)	75
Evaluative (neg.)	50
Evaluative (pos.)	43
Medical	83
Relational + Affiliative	95
Average	**63**

Accuracy assessment: UMAP after DR (K8)

Methodology: what have we done?

KNN Clustering

- Checked different numbers of clusters (K): here we present the results for 8 and 12 clusters
- Number of clusters: automated evaluations did not provide meaningful guidance
- Number of clusters should be comparable to existing classifications (Dixon (1977) and Biber et al. (2007))
- 8 clusters did not produce a fine-grained enough solution

Evaluation

- How did our results hold up against existing classifications (Biber et al. 2007)?
- Eye-balling classification (manual checks)
- Coherence metrics: How consistent were our results?
  - draw sample from each cluster to determine category type
  - generate new set of words based on category and check cluster allocation
  - generate Confusion Matrix and calculate coherence by comparing actual and predicted clustering

Discussion

Issues, Limitations, and Outlook

Discussion: Issues, Limitations and where from here?

What have we learned?

A semantic classification using usage-based data is possible, however there are still issues that require addressing:

- **Coherence**: some clusters proved to be less coherent or distinctive as would be desirable
- **Polysemy**: different senses of a type need to be addressed
- **PoS accuracy**: the current study pos-tagging was inaccurate as we relied on a data set of pre-existing word embeddings trained on news texts
- **Word embeddings**: we relied on a data set of pre-existing word embeddings trained on news texts
- **Evaluation**: our metrics are somewhat crude, but we are unsure how to assess the quality of our method (aside from reproducibility)

Discussion: Issues, Limitations and where from here?

What can we do?

- **Compile data and self-generate word embeddings**: we will generate our own word embeddings (using less but more diverse data) to avoid genre/text type bias and to allow the incorporation of polysemy (this will also dramatically improve pos accuracy)
- **Coherence**: once we have generated custom word embeddings, we can adapt our approach and model parameters
  - optimizing K in KNN clustering
  - Number of neighbors in UMAP
  - Try alternative classifiers (which potentially improves coherence)
- **Evaluation**: try out alternative evaluation methods (happy for input!)

Thank you very much!

Contact

m.schweinberger@uq.edu.au

Slides and resources

https://github.com/MartinSchweinberger/ICAME45

References

Dixon, R.M.W. (1977). Where have all the adjectives gone?. *Studies in Language*, *1*, 19-80.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (2007). *Grammar of Spoken and Written English*. Pearson Education Limited.

Chandrasekraran, D. & Mago, V. (2021). Evolution of Semantic Similarity – A Survey. *ACM Computing Surveys, 51(2),* Article 41. https://doi.org/10.1145/3440755

Haripse, S., Ranwez, S., Janaqi, S., Montmain, J. (2022). *Semantic Similarity from Natural Language and Ontology Analysis*. Springer Nature Switzerland AG

Hatrigan, J.A, Wong, M.A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Applied statistics, 28*, 100-108. doi:10.2307/2346830.

McInnes, L., Healy, J. and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv.org.* https://doi.org/10.48550/arXiv.1802.03426

Parker, R., Graff, D., Kong, J., Chen, K., Maeda, K. (2011). *English Gigaword Fifth Edition.* https://doi.org/10.35111/wk4f-qt80

Starmer, J. (2023, March 13). Word Embedding and Word2Vec, Clearly Explain!!!. https://www.youtube.com/watch?v=viZrOnJclY0

Starmer, J. (2018, May 24). StatQuest: K-Means clustering. https://www.youtube.com/watch?v=viZrOnJclY0

Rayson, P., Archer, D. & Piao, S. (2004). The UCREL semantic analysis system. *ResearchGate*.

Tagliamonte, S. (2008). So different and pretty cool! Recycling intensifiers in Toronto, Canada. *English Language and Linguistics, 12(2),* 361-394. doi:10.1017/S1360674308002669

Wijffels, J. (2023). *Udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe NLP Toolkit.* https://CRAN.R-project.org/package=udpipe.

Franti, P., Virmajoki, O., & Hautamaki, V. (2006). Fast agglomerative clustering using a k-nearest neighbor graph. IEEE transactions on pattern analysis and machine intelligence, 28(11), 1875-1881.

THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

Automated, Corpus- and Usage-Based Semantic Classification of Word Class using Word Embeddings

Martin Schweinberger and Chang-Hao (Howard) Luo

The University of Queensland

Contact

m.schweinberger@uq.edu.au

Slides and resources

https://github.com/MartinSchweinberger/ICAME45