Similar, but Different

Unsupervised Detection of Semantic Shifts in Diachronic Word Embeddings

Simon Preissner
s8siprei@stud.uni-saarland.de

Yuri Bizzoni
yuri.bizzoni@uni-saarland.de

Elke Teich
e.teich@mx.uni-saarland.de

What's the Matter?

Language is different across uses and across time.

Comparing word embeddings can inform on semantic shifts.

Especially diachronic scenarios involve subtle shifts.[1]

Humans often don't see them.

Detect semantic shifts
without presuppositions and expectations

Diachronic Data: the RSC

– Royal Society Corpus: scientific texts in British English

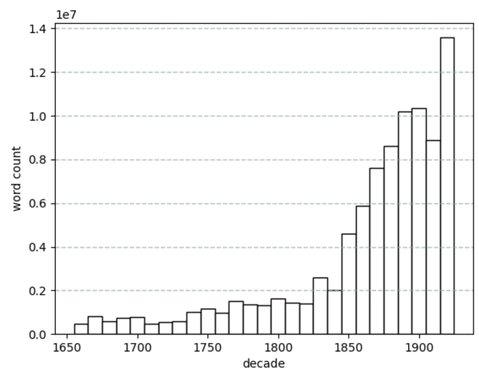
– total size: 91M (sizes per decade: 0.45M – 13.5M)

– time span: 1660-1930 [2]

– Word2Vec embeddings [3]
(structured skip-grams) [4]

– decade-wise embeddings,
each building on the previous one
(first decade: on the whole RSC)

– including per-decade token counts



Gromov-Wasserstein Optimal Transport [5] (GWOT)

OT finds the cheapest reallocation of mass for two point sets.[6]

GW enables comparisons across embedding spaces. [7]

Concepts have similar geometric constellations across spaces.[8]

GWOT looks at within-space distances
and matches points by their spatial 'role'

GWOT aims to align two spaces perfectly.

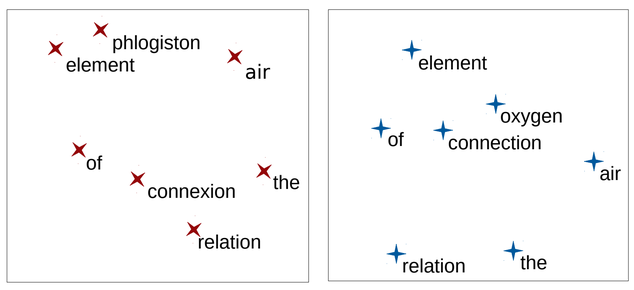
Aligning only the most stable concepts preserves the semantic shifts.

Shift Detection Methods

given: spaces **X** and **Y**, translation pairs **T**, projection matrix **P**

1. string mismatches among pairs $(v,w) \in T$
   confident mismatches can indicate large shifts
2. self-distance: $1 - cos(Px, x) : x \in X$
   operates only on the source space
3. cosine distance: $1 - cos(Px_v, y_w) : (v,w) \in T$
   and CSLS[9]
4. clustering of difference vectors: $y_w - Px_v : (v,w) \in T$
   detect systematic/thematic shifts and their directions

Imperfect Alignment – on Purpose

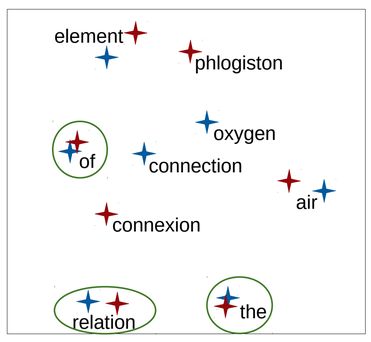


GWOT creates a **coupling:** a bidirectional table of translation likelihoods.

This leads to pairs of mutual best translations.

1. pick only the most likely (~stable) pairs as "**anchors**" for alignment.
   → small coupling (<200 words)
   → Procrustes' on the pairs' vectors
   → project X onto Y

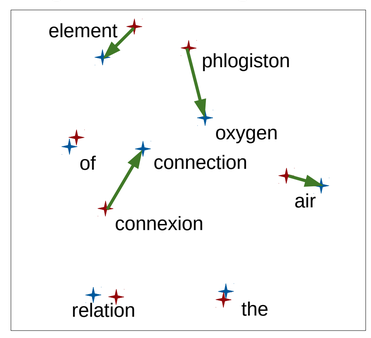
   [(of,of),
   (the,the),
   (relation,relation)]



align X with Y along anchors

2. this is seen as a bilingual task;
   we need a set of **translation pairs**.
   → large coupling (>5K words)
   → measure shift for these pairs

   [(air, air),
   (relation, relation),
   (element,element),
   (connexion, connection),
   (phlogiston, oxygen)]



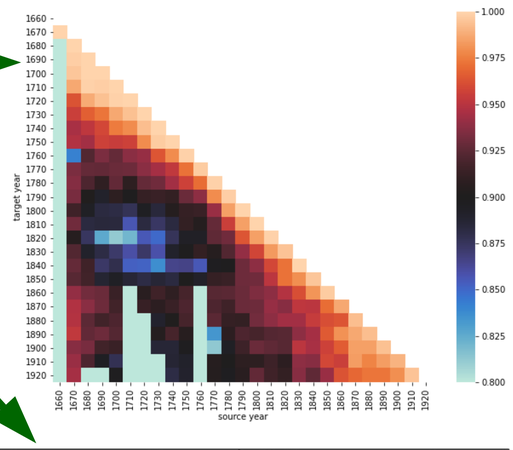
shifts between translation pairs

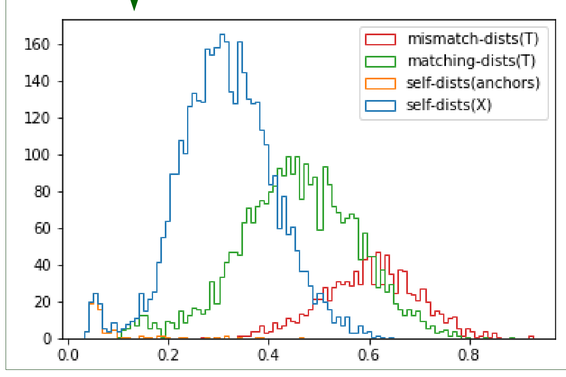
Preliminary Results

farther apart?
harder to align.

distances of translation pairs
and self-distances of their x

words with large shifts
from 1850 to 1900





distance	word	distance	word
0.8824	standards	0.7611	broke
0.8288	tons	0.7548	pendulum
0.7962	suspension	0.7528	vena
0.7803	extinct	0.751	chain
0.7793	equatorial	0.7504	glands
0.7734	stone	0.7489	inertia
0.772	s	0.7471	model
0.7697	ciliated	0.7433	geometry
0.7684	deposition	0.7433	metre
0.7615	diffuse	0.7417	arc

References

[1] Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. *CoRR, abs/1806.03537*. http://arxiv.org/abs/1806.03537
[2] Fischer, S., Knappen, J., Menzel, K., & Teich, E. (2020). *The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study*. 9.
[3] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv:1301.3781*.
[4] Ling, W., Dyer, C., Black, A. W., & Trancoso, I. (2015). Two/Too Simple Adaptations of Word2Vec for Syntax Problems. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1299–1304. https://doi.org/10.3115/v1/N15-1142
[5] Alvarez-Melis, D., & Jaakkola, T. (2018). Gromov-Wasserstein Alignment of Word Embedding Spaces. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1881–1890. https://doi.org/10.18653/v1/D18-1214
[6] Peyré, G., & Cuturi, M. (2019). Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, *11*(5–6), 355–607. https://doi.org/10.1561/2200000073
[7] Mémoli, F. (2011). Gromov–Wasserstein Distances and the Metric Approach to Object Matching. *Foundations of Computational Mathematics*, *11*(4), 417–487. https://doi.org/10.1007/s10208-011-9093-5
[8] Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting Similarities among Languages for Machine Translation. *ArXiv:1309.4168 [Cs]*. http://arxiv.org/abs/1309.4168
[9] Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2018). Word Translation Without Parallel Data. *ArXiv:1710.04087 [Cs]*. http://arxiv.org/abs/1710.04087

UNIVERSITÀ DEGLI STUDI DI TRENTO
UNIVERSITÄT DES SAARLANDES
LCT