Time for change: Evaluating models of semantic change without evaluation tasks

Haim Dubossarsky[1], Simon Hengchen[2], Nina Tahmasebi[3], & Dominik Schlechtweg[4]

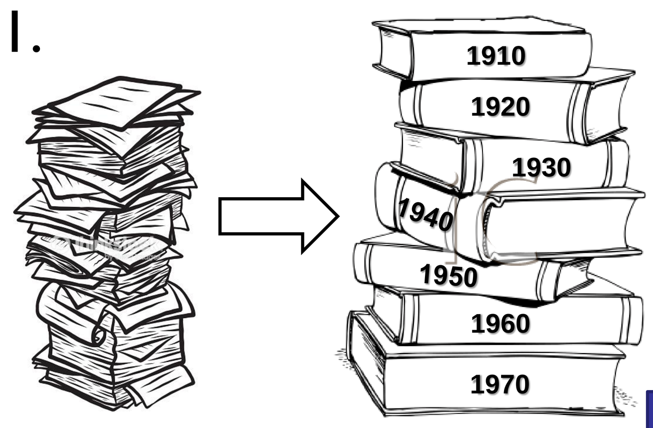
All authors contributed equally for this work, and the order was randomly assigned.

[1]Language Technology Lab, University of Cambridge; [2]COMHIS, University of Helsinki; [3]Department of Swedish, University of Gothenburg; [4]Institute for Natural Language Processing, University of Stuttgart

Noise factors in common pipeline for semantic change analysis

Split and align – two sources of noise

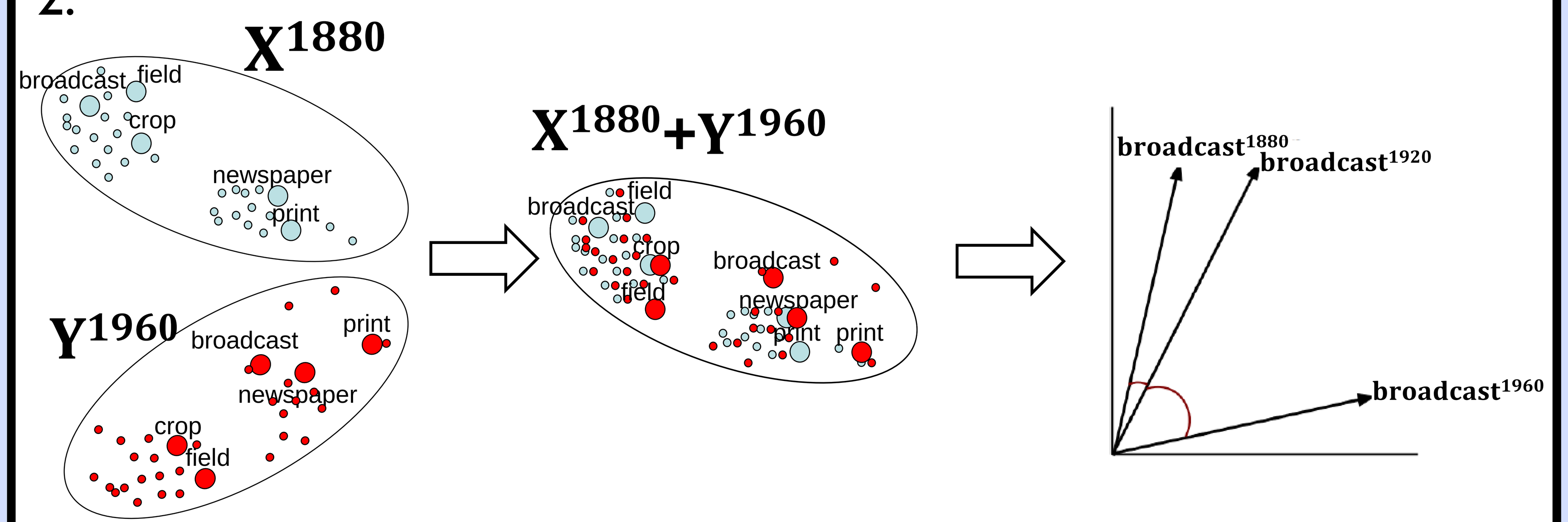
1.



- An original corpus C is split into sub-corpora time bins, $C_a, C_b, \ldots, C_n$.
- Embedding are trained on each bin separately.
- This "downsample" the words' frequency, as each embedding in based on smaller sample.

Embedding is noisier for lower frequency words.

2.



- Orthogonal Procrustes Analysis is computed between two embedding spaces:
$$W^* = argmin_W \|X^{1880}W - Y^{1960}\|$$
and applied to make the spaces **aligned** and comparable.
- Embeddings of the same word from two time bins are compared using cosine-similarity, which provide an estimate for lexical semantic change for that word.

Alignment is not perfect and introduces noise.

Temporal referencing[1,2]

Temporal referencing (TR) supports training on the original corpus, which circumvent the *split* and *align* steps and their **assumed** noise.

Example

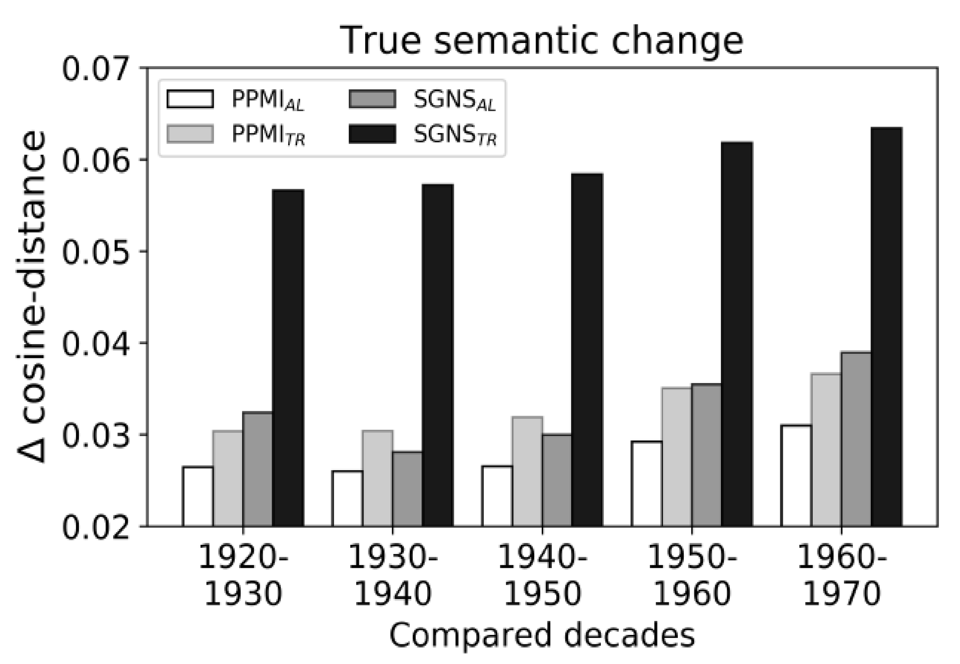
```
Silken cauliflowers sown broadcast1870 over the land.
The dramatic broadcast1970 stunned the nation.
```

Following comparisons would inform us about the assumed sources of noise.

Model		
$PPMI_{AL}$	Testing for separate noise from downsampling	Testing for separate noise from alignment
$PPMI_{TR}$		
$SGNS_{AL}$	Testing for combined noise from downsampling and alignment	
$SGNS_{TR}$		

Experiment 1 – TR is less noisy

Performance under a shuffled corpus provides an estimate for noise levels[3]. Comparison to the original corpus provides an estimate for true effect size.



Downsampling and alignment are two independent sources of noise.
Noise by alignment is much greater than by downsampling.

Reference list

[1]Alessio Ferrari, Beatrice Donati, and Stefania Gnesi. 2017. Detecting domain-specific ambiguities: an NLP approach based on wikipedia crawling and word embeddings. In IEEE, pages 393–399.

[2]Dominik Schlechtweg, Anna H¨atty, Marco del Tredici, and Sabine Schulte im Walde. 2019. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In Proceedings of ACL.

[3]Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In EMNLP 2017, pages 1136–1145.

[4]Nina Tahmasebi and Thomas Risse. 2017. Word sense change testset, 10.5281

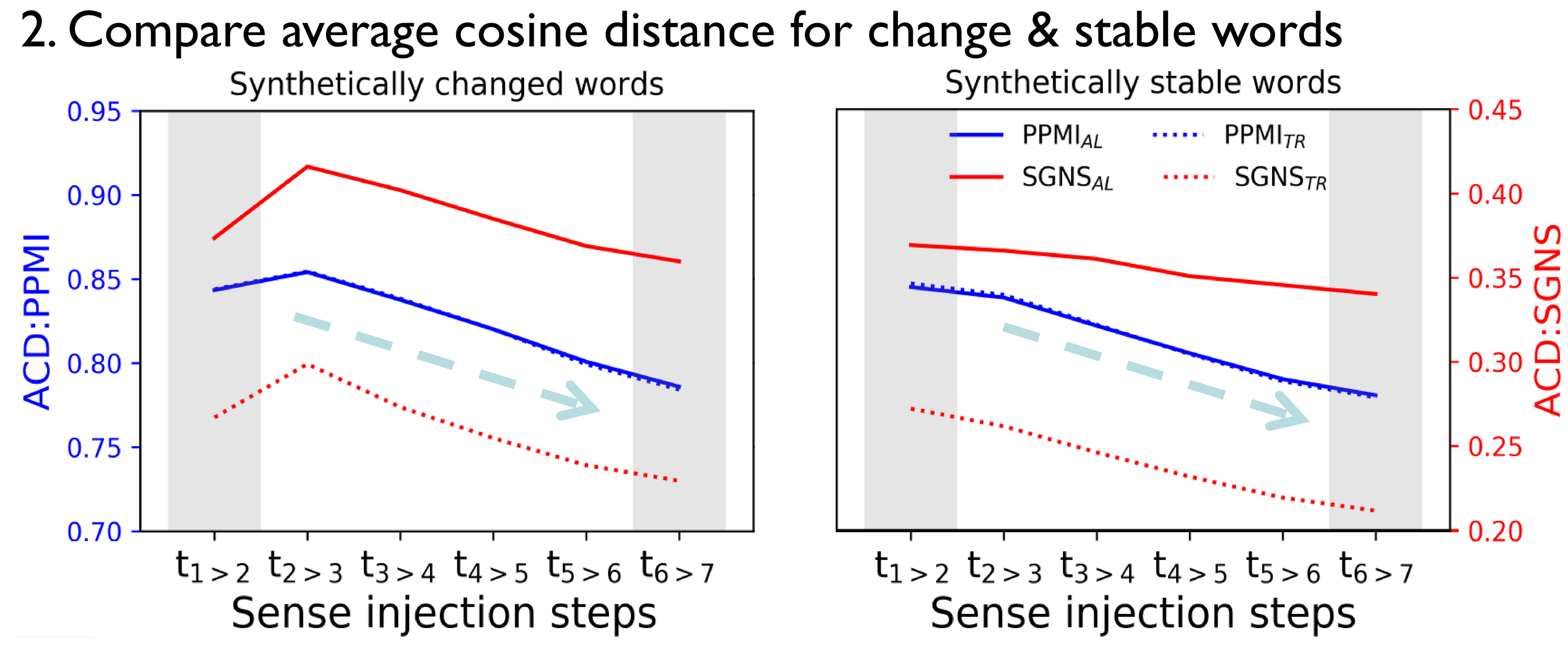
Experiment 2 – TR is better in detecting synthetic change

1. Injecting synthetic semantic change into a corpus (for 356 words)

	Original text		Text with injected change	Change ratio
t_1	A wedding ring An arm bracelet	➔	A wedding ring	[100%] [0%]
t_2	A wedding ring An arm bracelet	➔	A wedding ring	[100%] [0%]
t_3	A wedding ring An arm bracelet	➔ ➔	A wedding ring An arm ring	[100%] [25%]
t_4	A wedding ring An arm bracelet	➔ ➔	A wedding ring An arm ring	[100%] [50%]
t_5	A wedding ring An arm bracelet	➔ ➔	A wedding ring An arm ring	[100%] [75%]
t_6	A wedding ring An arm bracelet	➔ ➔	A wedding ring An arm ring	[100%] [100%]
t_7	A wedding ring An arm bracelet	➔ ➔	A wedding ring An arm ring	[100%] [100%]

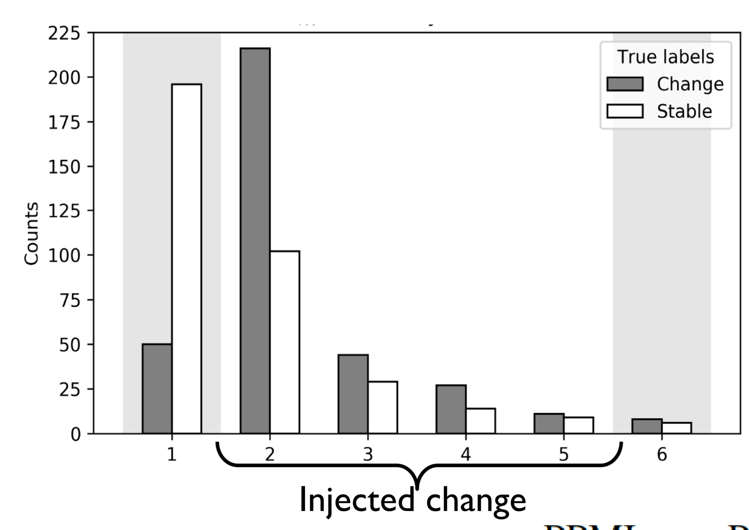
* Additional 356 stable control words match the frequency increase
** Steps without injection are shaded.

2. Compare average cosine distance for change & stable words



Synthetic change validated, change words are markedly different than stable words for all models.

3. Synthetic semantic change as a classification task




```
Train naïve classifier

if 2=<peak_position=<5:
    semantic_change = True
else:
    semantic_change = False
```

	$PPMI_{AL}$	$PPMI_{TR}$	$SGNS_{AL}$	$SGNS_{TR}$
Stable	0.52	0.54	0.37	**0.57**
Unrelated	0.83	0.83	0.86	**0.91**
Related	0.73	0.73	0.78	**0.78**
Mean acc.	0.65	0.66	0.59	**0.70**
F1-score	0.69	0.69	0.67	**0.74**

All models perform better than chance in detecting synthetic semantic change.
TR has the best performance!

Experiment 3 – TR is better in detecting attested change[4]

	SGNS		PPMI	
	Align	TR	Align	TR
Change	0.47	0.31	0.86	0.86
Stable	0.34	0.21	0.71	0.73
DIFF	38%	50%	20%	17%

TR shows the largest increase between change and stable words (13 change, 19 stable).

Conclusions

1. Downsampling and alignment each introduces a separate source of noise.
2. TR allows to train embedding not exposed to any of these two noises.
3. TR is better at detecting synthesis as well as attested semantic change.
4. TR provides a less nosier model as well as better detection for semantic change.