# Sequential Modelling of the Evolution of Word Representations for Semantic Change Detection

Adam Tsakalidis & Maria Liakata

EMNLP 2020

Queen Mary University of London
The Alan Turing Institute
WARWICK THE UNIVERSITY OF WARWICK

# Lexical Semantic Change

*Ah, look at those **blackberries**, aren't they beautiful?*

*Where? I can't see them...*

*There! Where this little guy is **tweeting**!*

*Huh...?*

<u>Task</u>: Identify words that change their meaning over time (Tahmasebi et al., 2018)

Applications: historical linguistics, evolution of communities, cultural shifts...

# Related Work

Diachronic representations:

topics (Frermann & Lapata, 2016; Perrone et al., 2019)
graphs (Mitra et al., 2014)
**neural** (Hamilton et al., 2016; Shoemark et al., 2019; Schlechtweg et al., 2019)

Common practice (Hamilton et al., 2016):

- Learn word representations in two distinct time periods
- Align them via Orthogonal Procrustes (Schönemann, 1966)
- Measure cosine similarity

+Pros: Highly effective, fast

-Cons: Temporality? Non-linear?

# Related Work & Contributions

Approaches taking time into consideration:

- Rely on linear transformations (Kulkarni et al., 2015; Shoemark et al., 2019)
- Focus on *word representation* (Kim et al., 2014; Dubossarsky et al., 2019)

Our contributions:

- Work with *"any"* pre-trained word representations over time
- Non-linear, sequential models for semantic change
- Evaluate models in a sequential dataset; compare against strong baselines

# Proposed Method

Input: Pre-trained word vectors in T time periods:

$$[W_0, W_1, ..., W_{i-1}, W_i, W_{i+1}, ..., W_{T-1}]$$

Goal: Learn how the vectors evolve over time.

Semantic change: Words whose sequence is hard to predict

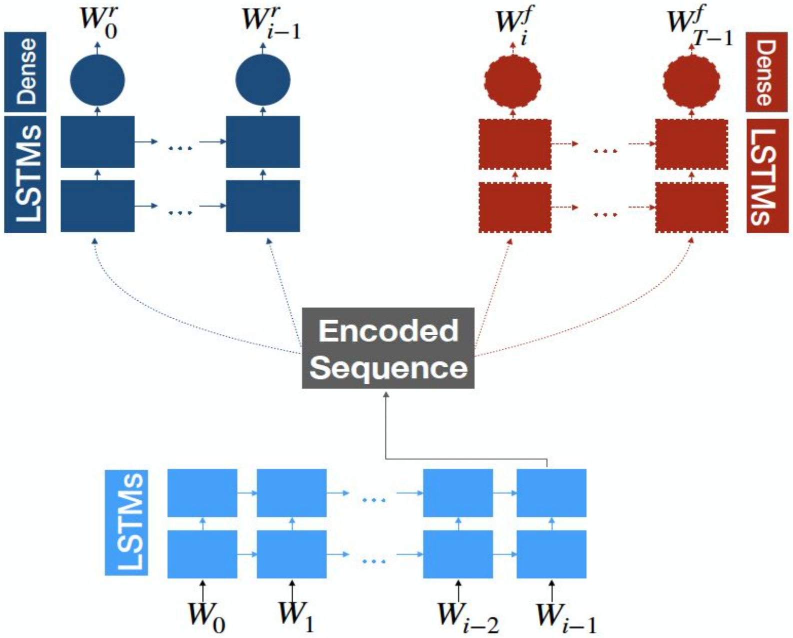
# Proposed Method

<u>Input</u>: Pre-trained word vectors in T time periods:

$$[W_0, W_1, ..., W_{i-1}, W_i, W_{i+1}, ..., W_{T-1}]$$

<u>Goal</u>: Learn how the vectors evolve over time.

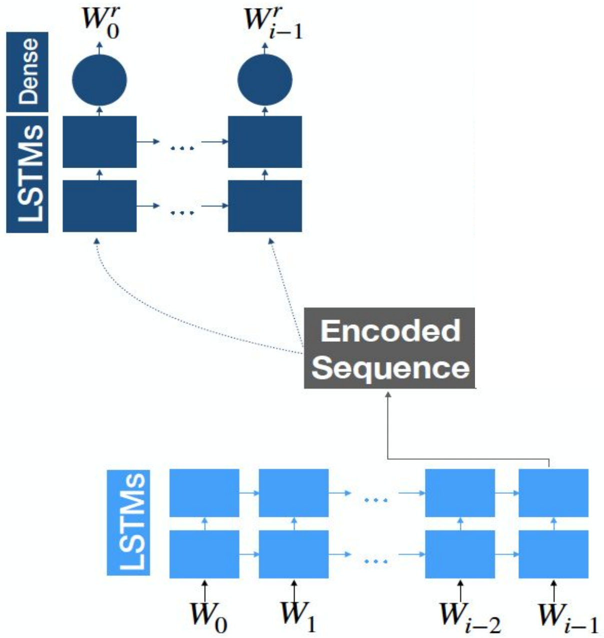
<u>How</u>: (a) autoencoder
(b) future prediction
(c) multi-task



## Proposed Method

Reconstruct input sequence of word vectors through time

$$L_r = \frac{1}{i}\sum_{j=0}^{i-1} MSE(W_j, W_j^r)$$




Input: Pre-trained word vectors in T time periods:

$$\boxed{[W_0, W_1, ..., W_{i-1}, W_i, W_{i+1}, ..., W_{T-1}]}$$

Goal: Learn how the vectors evolve over time.

How: (a) autoencoder $seq2seq_r$
(b) future prediction
(c) multi-task

# Proposed Method

Input: Pre-trained word vectors in T time periods:

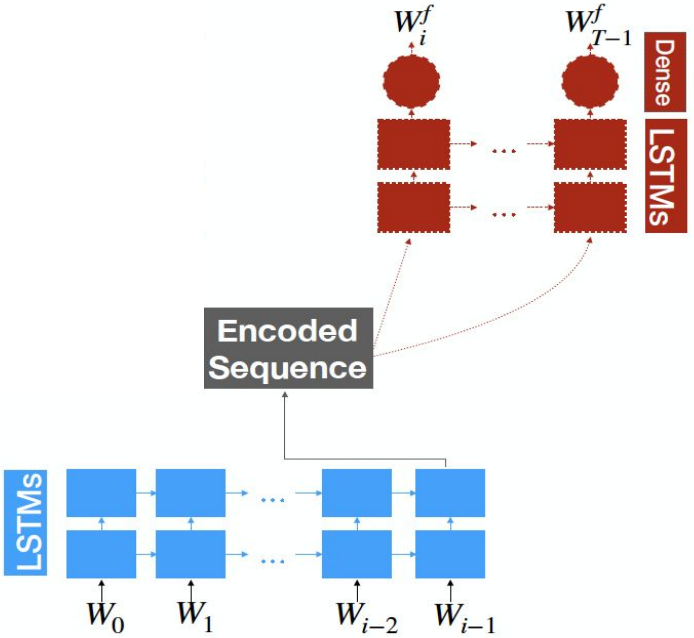
$$\boxed{[W_0,} W_1, ..., W_{i-1}, W_i, W_{i+1}, ..., W_{T-1}]$$

Goal: Learn how the vectors evolve over time.

How:
(a) autoencoder $seq2seq_r$
**(b) future prediction** $seq2seq_f$
(c) multi-task

Predict future sequence of word vectors through time

$$L_f = \frac{1}{T-i}\sum_{j=i}^{T-1} MSE(W_j, W_j^f)$$



# Proposed Method

Input: Pre-trained word vectors in T time periods:

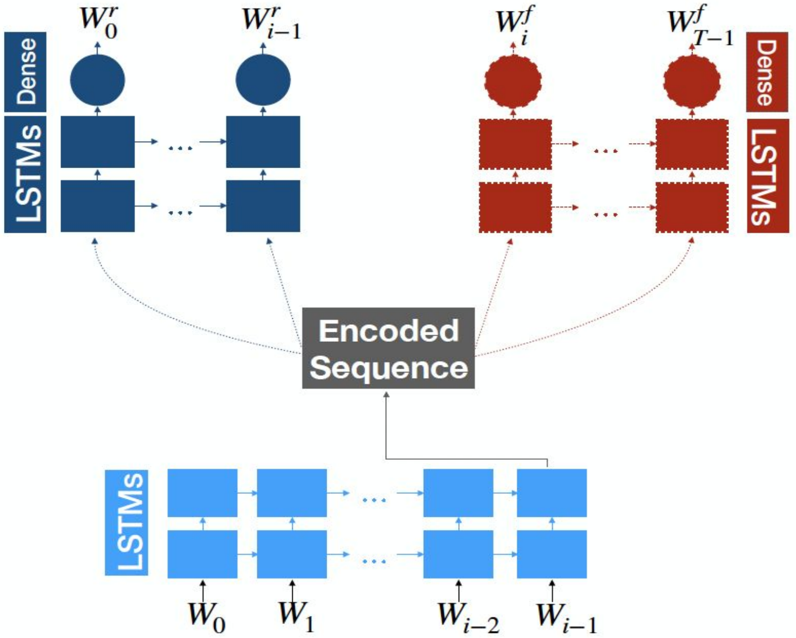
$$[\boxed{W_0, W_1, ..., W_{i-1}}, W_i, W_{i+1}, ..., W_{T-1}]$$

Goal: Learn how the vectors evolve over time.

How:
(a) autoencoder $seq2seq_r$
(b) future prediction $seq2seq_f$
(c) multi-task $seq2seq_{rf}$

Reconstruct past & predict future sequences

$$L_r = \frac{1}{i}\sum_{j=0}^{i-1} MSE(W_j, W_j^r) \qquad L_f = \frac{1}{T-i}\sum_{j=i}^{T-1} MSE(W_j, W_j^f)$$



# Dataset

UK Web Archive (Tsakalidis et al., 2019)

Size: 47.8K words

Time period: 2000-2013 - each year corresponds to a timestep in our modelling

Vectors: 100-dim (Mikolov et al., 2013), trained on each year independently

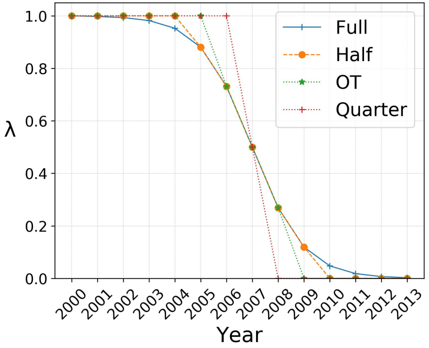
Split: 80/20 (train/test)

# Proof of Concept (Experiments)

Synthetic Semantic Shift: force 5% of word vectors in the test set ($w^{(\alpha)}$) to shift their meaning towards $w^{(\beta)}$ over time:

$$\boxed{w_t^{*(\alpha)} = \lambda_t w_t^{(\alpha)} + (1 - \lambda_t) w_t^{(\beta)}}$$

## Setting λ:




## Selecting $w^{(\beta)}$:

Condition: $c - 0.1 < cos(w_0^{(\alpha)}, w_0^{(\beta)}) \leq c$

Note: high *c* value => challenging! $w^{(\alpha)}$ moves towards a more similar $w^{(\beta)}$ - lower level of change

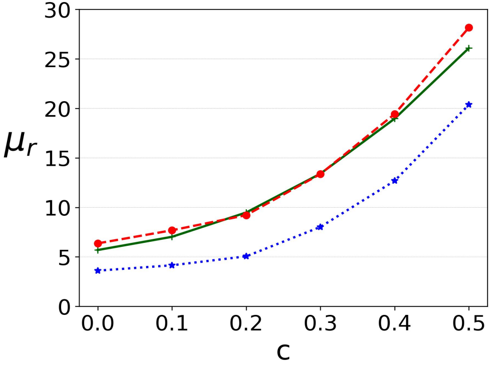
# Proof of Concept (Results)

Evaluation: Rank words in test set based on their average *cosDist*

**High *cosDist* => model failure => semantic change**

Metric: average rank of the semantically shifted words ( $\mu_r$)

**Full**




Note: high *c* value => challenging!

$w^{(\alpha)}$ moves towards a more similar $w^{(\beta)}$ - lower level of change

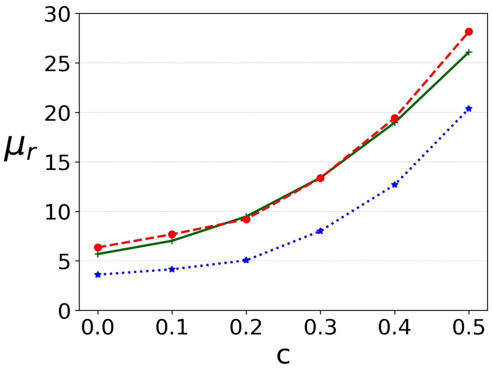
# Proof of Concept (Results)

Evaluation: Rank words in test set based on their average *cosDist*

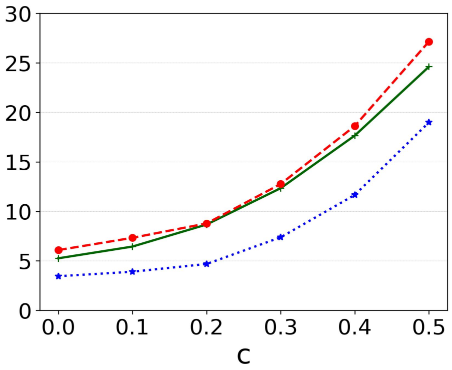
**High *cosDist* => model failure => semantic change**

Metric: average rank of the semantically shifted words ($\mu_r$)

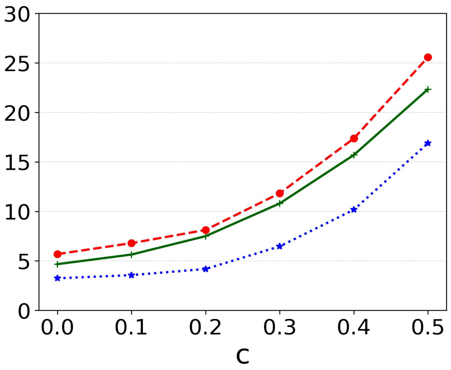
## Full



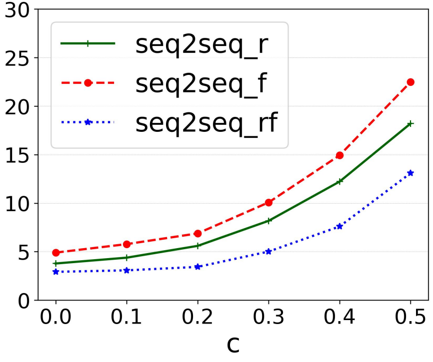

## Half




## OT




## Quarter



# Experiments & Comparisons

Data: 47.8K words - 65 words with altered meaning (Oxford English Dictionary)
Goal: Better rank for the 65 words with altered lexical semantics

Metrics: $\mu_r$ (lower is better)
$rec@k$ (higher is better)

Baselines: Models from prior work & alterations of our models

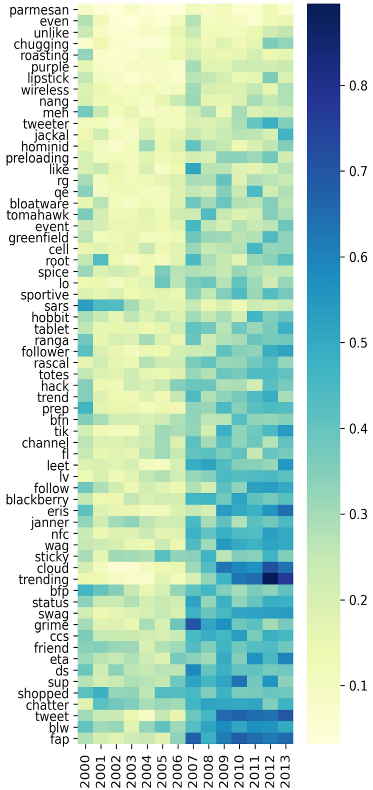
# Results

| | | μ_r | rec@5 | rec@10 | rec@50 |
|---|---|---|---|---|---|
| Prior Work | PROCR | 30.63 | 18.46 | 27.69 | 78.46 |
| | PROCR_k | 31.01 | 21.54 | 27.69 | 75.38 |
| | PROCR_kt | 31.91 | 20.00 | 27.69 | 70.77 |
| | RF | 30.01 | 10.77 | 21.54 | 78.46 |
| | LSTM_r | 27.87 | 12.31 | 29.23 | 80.00 |
| | LSTM_f | 28.62 | 16.92 | 32.31 | 76.92 |
| | GT_c | 47.87 | 7.69 | 16.92 | 52.31 |
| | GT_β | 38.09 | 13.85 | 24.62 | 66.15 |
| | PROCR_* | 25.01 | 21.54 | 32.31 | 81.54 |
| Ours | seq2seq_r | 24.75 | 21.54 | 38.46 | **84.62** |
| | seq2seq_f | **23.86** | 26.15 | **46.15** | **84.62** |
| | seq2seq_rf | 24.28 | **29.23** | 36.92 | **84.62** |

# Results

| | | μ_r | rec@5 | rec@10 | rec@50 |
|---|---|---|---|---|---|
| Prior Work | **PROCR** | 30.63 | 18.46 | 27.69 | 78.46 |
| | **PROCR_k** | 31.01 | 21.54 | 27.69 | 75.38 |
| | **PROCR_kt** | 31.91 | 20.00 | 27.69 | 70.77 |
| | **RF** | 30.01 | 10.77 | 21.54 | 78.46 |
| | **LSTM_r** | 27.87 | 12.31 | 29.23 | 80.00 |
| | **LSTM_f** | 28.62 | 16.92 | 32.31 | 76.92 |
| | **GT_c** | 47.87 | 7.69 | 16.92 | 52.31 |
| | **GT_β** | 38.09 | 13.85 | 24.62 | 66.15 |
| | **PROCR_*** | 25.01 | 21.54 | 32.31 | 81.54 |
| Ours | **seq2seq_r** | 24.75 | 21.54 | 38.46 | **84.62** |
| | **seq2seq_f** | **23.86** | 26.15 | **46.15** | **84.62** |
| | **seq2seq_rf** | 24.28 | **29.23** | 36.92 | **84.62** |

# Results




tweet, eta, friend...

parmesan, even, unlike

Future: Contextual word vectors (Giulianelli et al., 2020)

| | | μ_r | rec@5 | rec@10 | rec@50 |
|---|---|---|---|---|---|
| Prior Work | PROCR | 30.63 | 18.46 | 27.69 | 78.46 |
| | PROCR_k | 31.01 | 21.54 | 27.69 | 75.38 |
| | PROCR_kt | 31.91 | 20.00 | 27.69 | 70.77 |
| | RF | 30.01 | 10.77 | 21.54 | 78.46 |
| | LSTM_r | 27.87 | 12.31 | 29.23 | 80.00 |
| | LSTM_f | 28.62 | 16.92 | 32.31 | 76.92 |
| | GT_c | 47.87 | 7.69 | 16.92 | 52.31 |
| | GT_β | 38.09 | 13.85 | 24.62 | 66.15 |
| | PROCR_* | 25.01 | 21.54 | 32.31 | 81.54 |
| Ours | seq2seq_r | 24.75 | 21.54 | 38.46 | **84.62** |
| | seq2seq_f | **23.86** | 26.15 | **46.15** | **84.62** |
| | seq2seq_rf | 24.28 | **29.23** | 36.92 | **84.62** |

# Conclusion

3 variants of non-linear, sequential models for Semantic Change Detection

Experiments with synthetic/real data; comparison against strong baselines

Future Work:

- Contextual word representations (Devlin et al., 2018)
- Different languages/durations
- Anomaly detection approaches
- Different model architectures

# References


[1] Devlin et al., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.
[2] Dubossarsky et al., 2019. Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change. In *ACL*.
[3] Frermann & Lapata, 2016. A Bayesian Model of Diachronic Meaning Change. In *Computational Linguistics*.
[4] Giulianelli et al., 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. In *ACL*.
[5] Hamilton et al., 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *ACL*.
[6] Kim et al., 2014. Temporal Analysis of Language through Neural Language Models. In *Workshop on Language Technologies and Computational Social Science*.
[7] Kulkarni et al., 2015. Statistically Significant Detection of Linguistic Change. In *WWW*.
[8] Mikolov et al., 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*.
[9] Mitra et al., 2014. That's Sick Dude!: Automatic Identification of Word Sense Change across Different Timescales. In *ACL*.
[10] Perrone et al., 2019. GASC: Genre-aware Semantic Change for Ancient Greek. In *International Workshop on Computational Approaches to Historical Language Change*.
[11] Schlechtweg et al., 2019. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *ACL*.
[12] Schönemann, 1966. A Generalized Solution of the Orthogonal Procrustes Problem. In *Psychometrika*.
[13] Shoemark et al., 2019. Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings. In *EMNLP*.
[14] Tahmasebi et al., 2018. Survey of Computational Approaches to Lexical Semantic Change. In *ArXiv*.
[15] Tsakalidis et al., 2019. Mining the UK Web Archive for Semantic Change Detection. In *RANLP*.

# Thank you!

atsakalidis@turing.ac.uk
mliakata@turing.ac.uk

@adtsakal
@xrysoflhs

**Acknowledgements**: This work was supported by The Alan Turing Institute (grant EP/N510129/1) and by a Turing AI Fellowship to M. Liakata, funded by the Department of Business, Energy & Industrial Strategy.