What are token vectors and their uses?

Sandeep Soni

09/12/2024

Administravia

- Problem Set 1
- Reading Response 1
- Teams

Problem Set

- We'll give you some starter code and a some task descriptions.
- Most tasks require you to just fill in some code; some would require you to explore and be creative.
- You'll submit your code (typically a colab notebook) and a 2 page document that describes the task, methodology and key findings.
- Your report doesn't have to be too wordy. Instead, make good use of tables or figures.
- Individual assignments; no extensive use of Generative AI

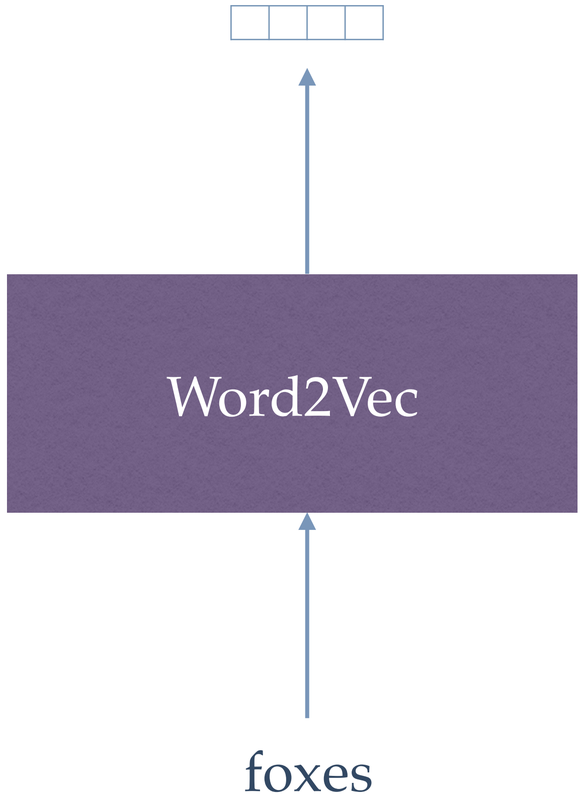
Reading response

- We'll ask you some multiple choice and free-form response questions.
- We're expecting crisp responses with adequate explanation.
- If you've been reading the required readings, the response should take a few hours (2-3) atmost.
- Individual assignments; No extensive use of Generative AI

Teams

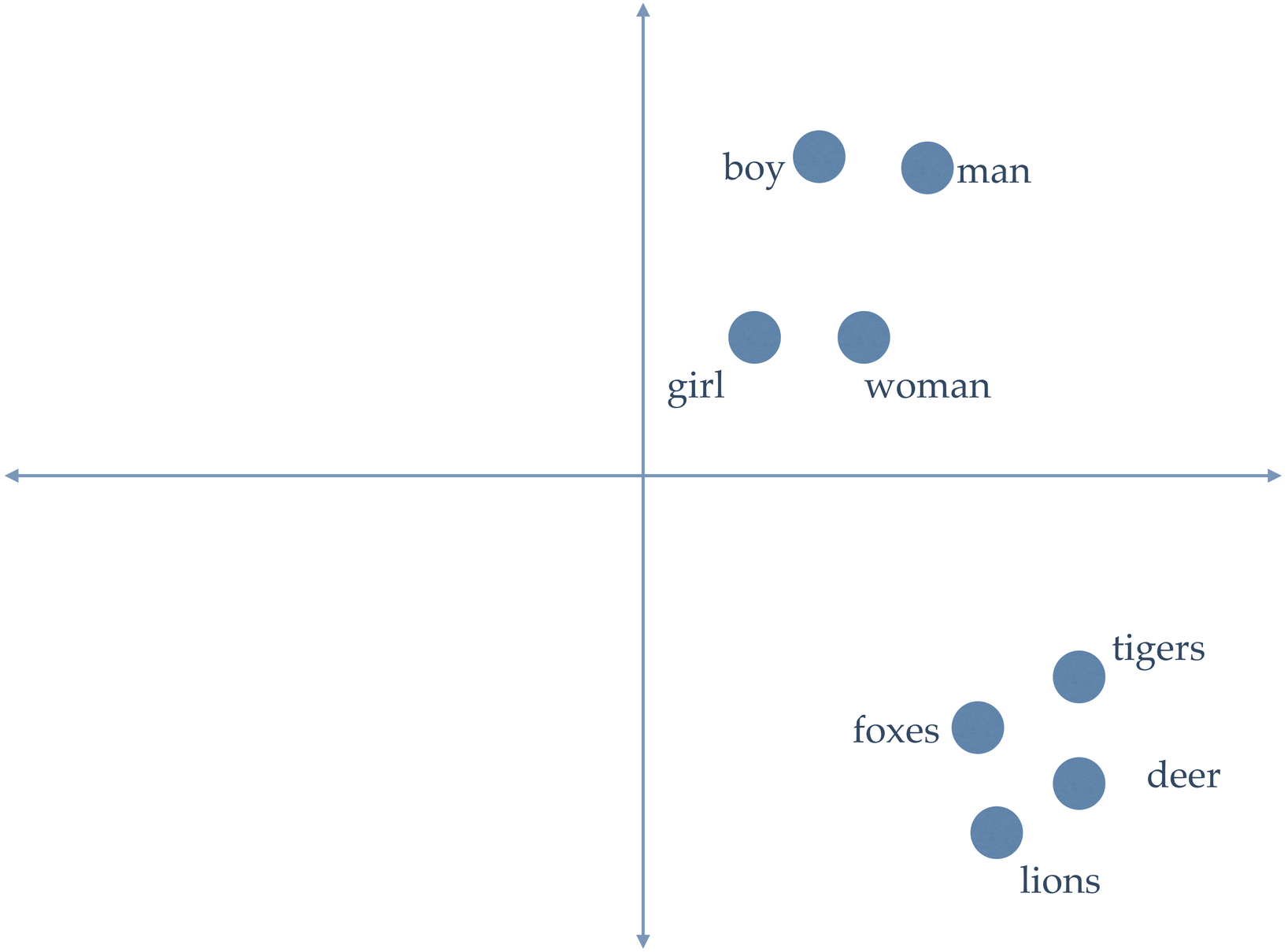
- Team project should be 3-4 people
- Please begin forming groups
- One person from each group should submit a list of members in their team, their email addresses and which year they are.

Word2vec abstraction



Input: word

Output:embedding



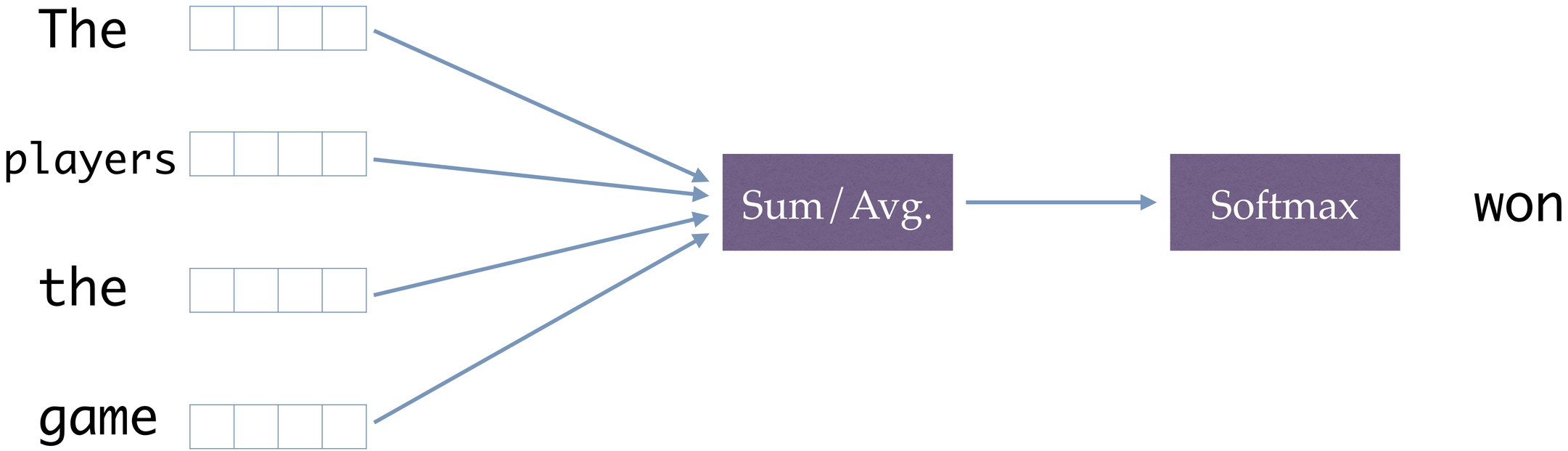
Semantic similarity transfers into geometric similarity

The players _____ the game

Fill in the blank by using the surrounding context

$$P(w_t \mid \neg w_t)$$

Word2Vec



Corpus

The players won the game
The coach praised the players
The players were victorious
…
The players won

Corpus

The players won the game
The coach praised the players
The players were victorious
…
The players won

How to measure the similarity between sentences?

Corpus

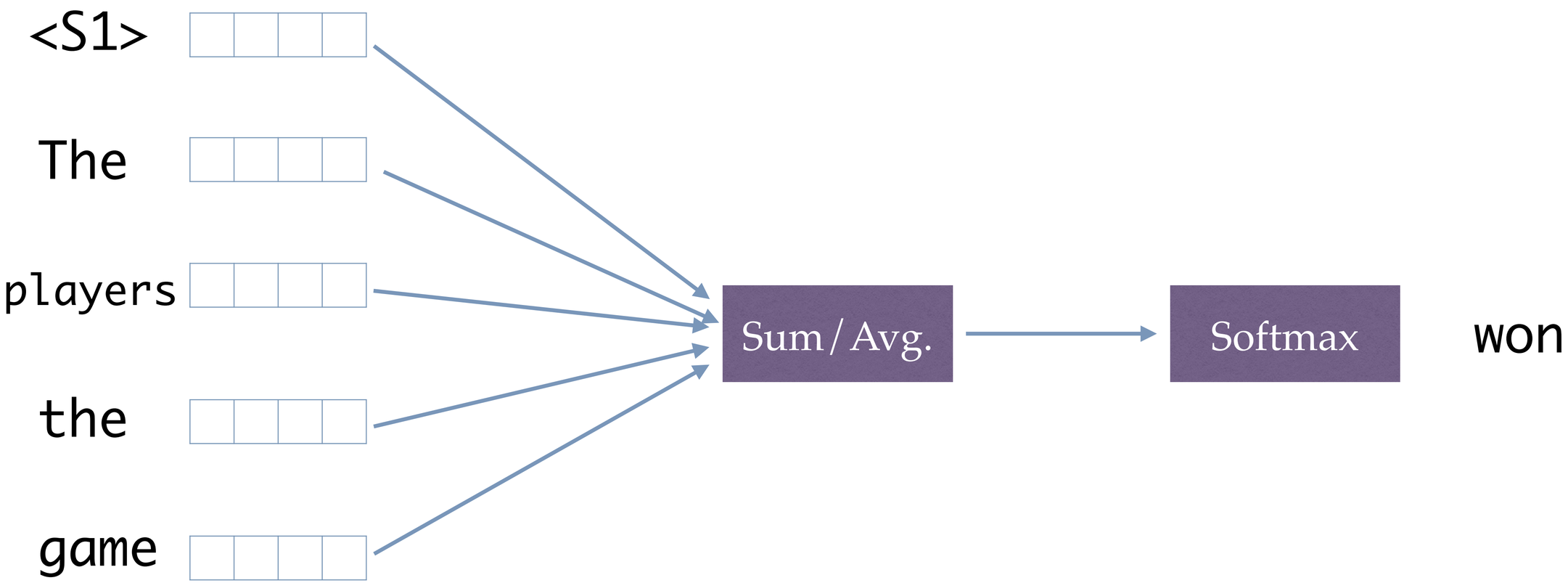
S1	The players won the game
S2	The coach praised the players
S3	The players were victorious
…	…
Sn	The players won

Corpus

S1	The players won the game
S2	The coach praised the players
S3	The players were victorious
…	…
Sn	The players won

<S1> The players won the game
<S2> The coach praised the players
<S3> The players were victorious
…
<Sn> The players won

Doc2Vec



Le and Mikolov (2014) Distributed Representations of Sentences and Documents

Corpus

S1	The players won the game
S2	The coach praised the players
S3	The players were victorious
…	…
Sn	The players won

How to measure the similarity between sentences?

Corpus

A1	The players won the game
A2	The coach praised the players
A3	The players were victorious
…	…
An	The players won

How to measure the similarity between authors?

Corpus

N1	The players won the game
N2	The coach praised the players
N3	The players were victorious
…	…
Nn	The players won

How to measure the similarity between newspapers?

Corpus

N1	The players won the game
N2	The coach praised the players
N3	The players were victorious
…	…
Nn	The players won

Data in the form of (tagged) sequences

How to use Word2Vec to measure the similarity of a song without knowing its lyrics?

I stay out too late
Got nothing in my brain
That's what people say, mm-mm
That's what people say, mm-mm
...

Tryna rain, tryna rain on the thunder
Tell the storm I'm new
I'm a wall, come and march on the regular
Painting white flags blue
...

I stay out too late
Got nothing in my brain
That's what people say, mm-mm
That's what people say, mm-mm
...

ts_song1

Tryna rain, tryna rain on the thunder
Tell the storm I'm new
I'm a wall, come and march on the regular
Painting white flags blue
...

b_song1

Corpus

playlist1	ts_song1 b_song4 e_song3 …
playlist2	ts_song5 b_song3 j_song1 …
playlist3	sd_song1 ts_song1 e_song5 …
…	…
playlistn	ts_song1 b_song1 …

How to measure the similarity between songs?

What is a limitation of word2vec?

- The _____ ran at us
- The _____ won the game

- `The foxes ran at us`
- `The foxes won the game`

Ideally, we want to give a different vector representation to every instance of a word

Types & Tokens

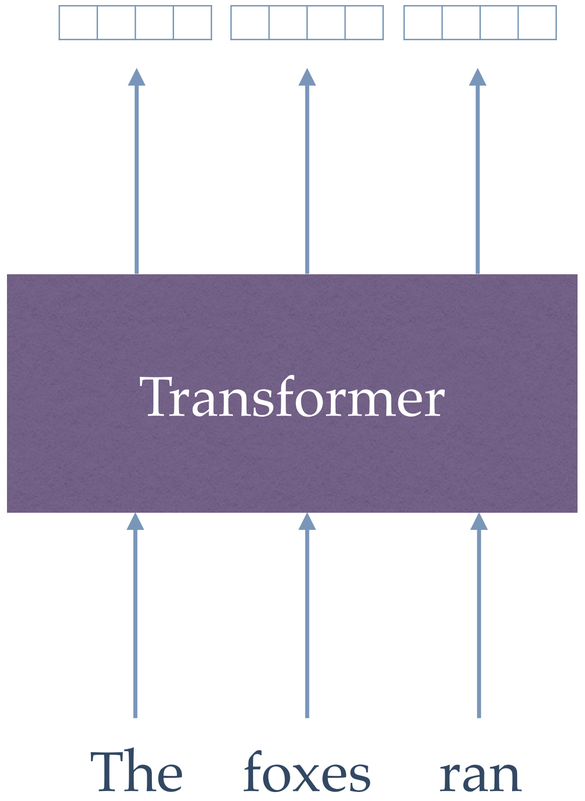
- Type: foxes
- Tokens:
  - The foxes ran at us

1.2	-0.1	0.7	-0.5

  - The foxes won the game

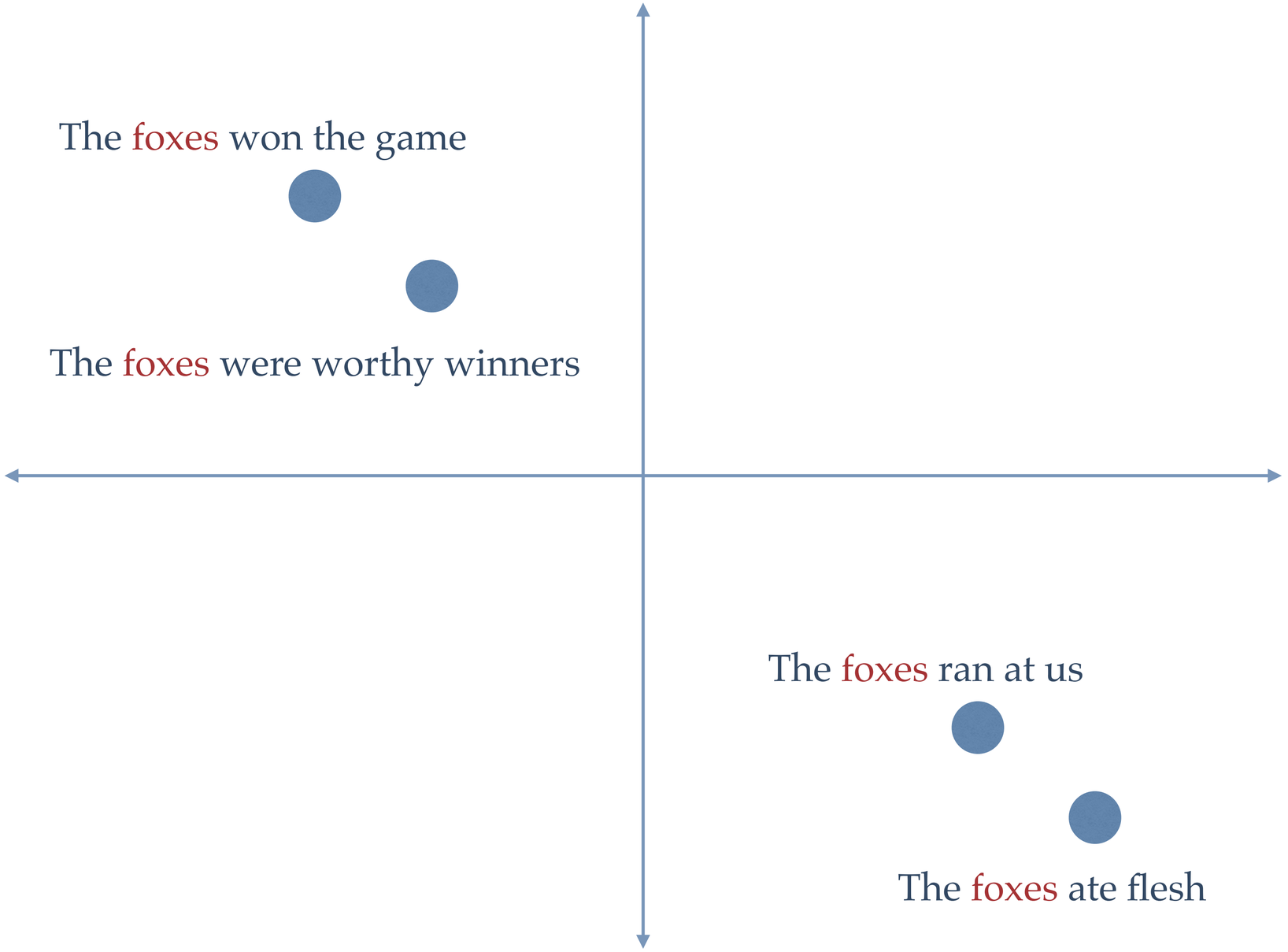
-0.4	0.6	0.8	0.4

Sequence Models: Abstraction



Input: Sequence of words

Output: Sequence of embeddings



Contextual similarity transfers into geometric similarity

Contextualized word Vectors

Transform static embedding to an embedding sensitive to the local context

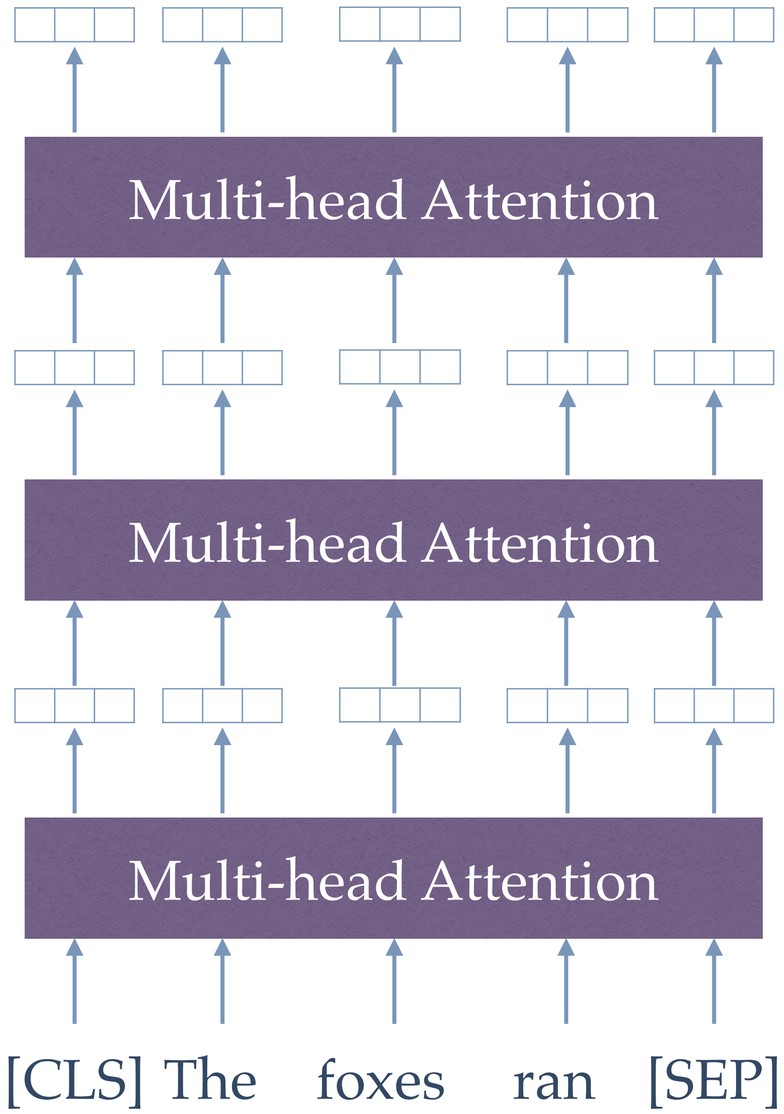
We further want these embeddings to be learned from training data and which can be updated for any task

BERT

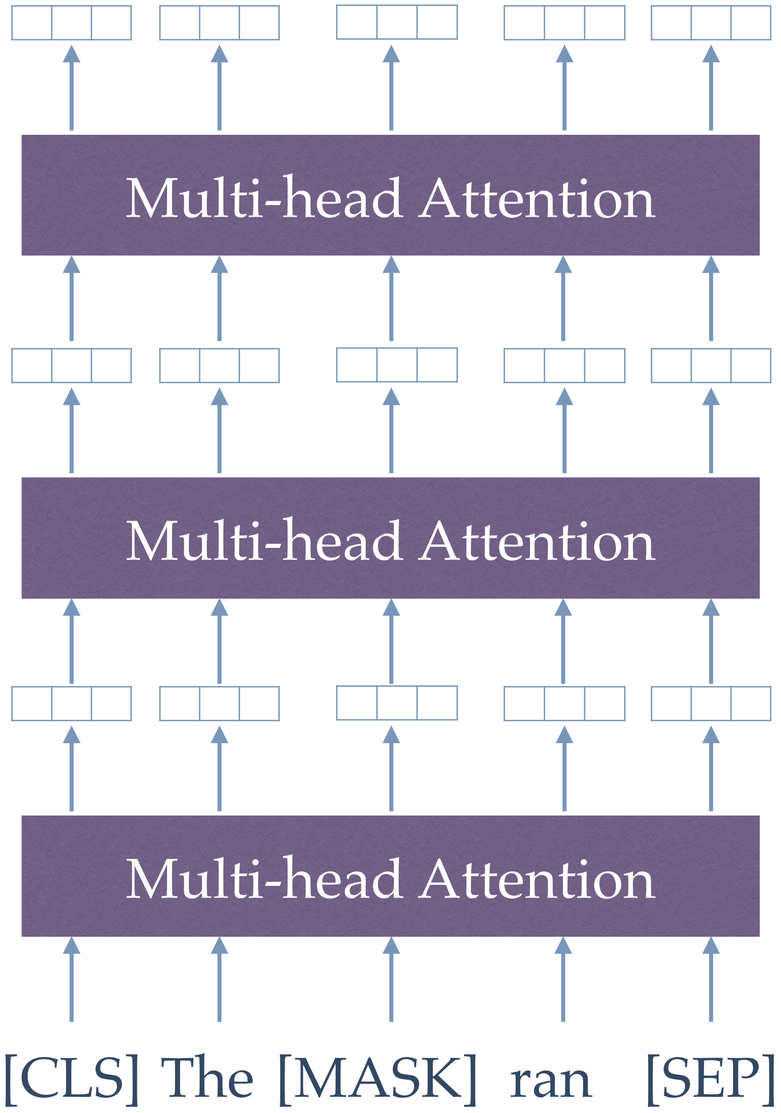
- Transformer based model that predicts masked word based on bidirectional context and next sentence prediction
- Multiple transformer blocks that take sequence of input vectors and give sequence of output vectors

Special Tokens

- Every sentence is appended with a special token [CLS] in the beginning and [SEP] at the end
- Embeddings for [CLS] and [SEP] tokens are also learned and can be used as vector representations of the entire sequence



Many such transformer blocks staked together make the BERT model



At the time of training, we try to predict the original token in place of MASK token

BERT

- Deep networks (12 layers for BERT base, 24 for BERT large)
- Token representations are high dimensional (768 dims for BERT base, 1024 for BERT large)
- Pretrained on large amounts of English text such as Wikipedia (2.5B words) and BooksCorpus (800M words)

Wordpiece

- Tokens are called wordpieces which allows to limit the vocabulary size and share subword information

this	this
grow	grow
growing	grow + #ing

“What can we do with contextual embeddings?”

Task performance

Plugging in the contextual embeddings can improve performance on linguistic tasks

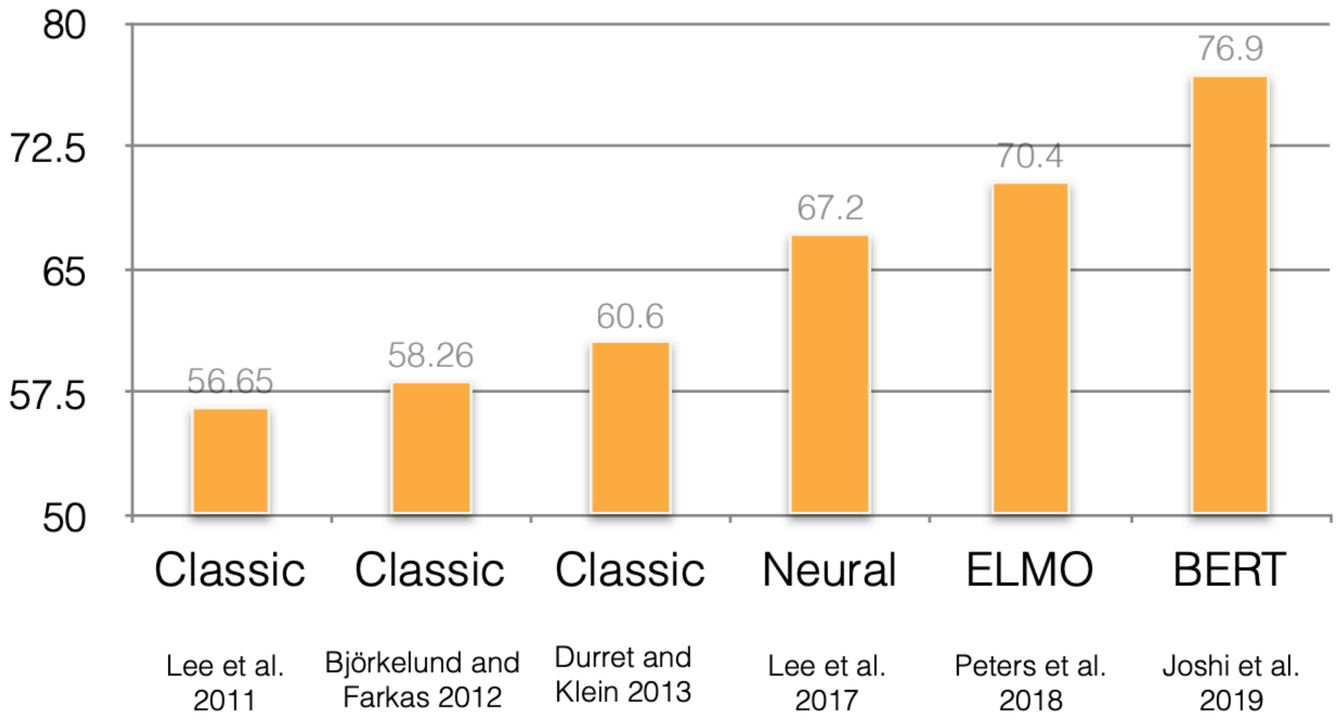
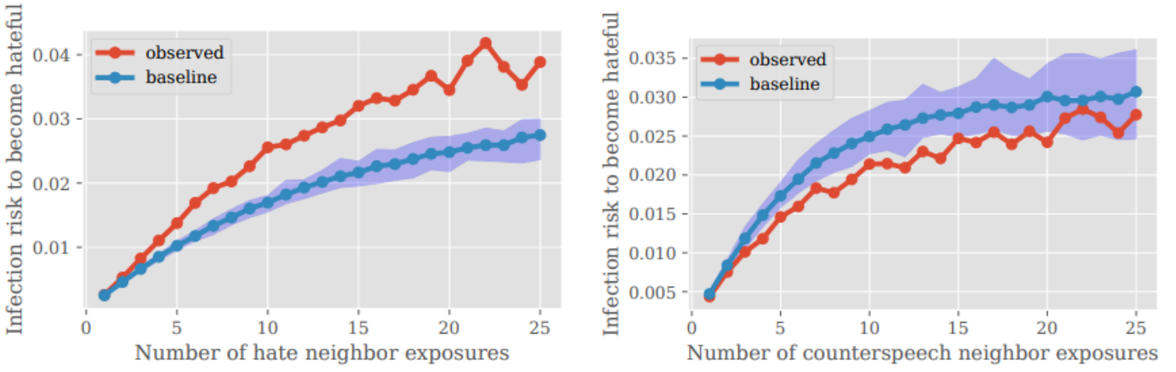


Figure taken from David Bamman's slide

Feature set	Precision	Recall	F1 score
Anti-Asian hate tweet detection			
Linguistic	0.541	0.233	0.323
Hashtag	0.100	0.002	0.005
BERT	0.765	0.760	0.762
Counterspeech tweet detection			
Linguistic	0.483	0.189	0.267
Hashtag	0.800	0.029	0.056
BERT	0.839	0.868	0.853
Neutral tweet detection			
Linguistic	0.632	0.891	0.739
Hashtag	0.591	0.999	0.743
BERT	0.886	0.874	0.880

TABLE III: Tweet classification performance of different feature sets with a neural network classifier. The BERT model has the best classification performance in all three tasks.

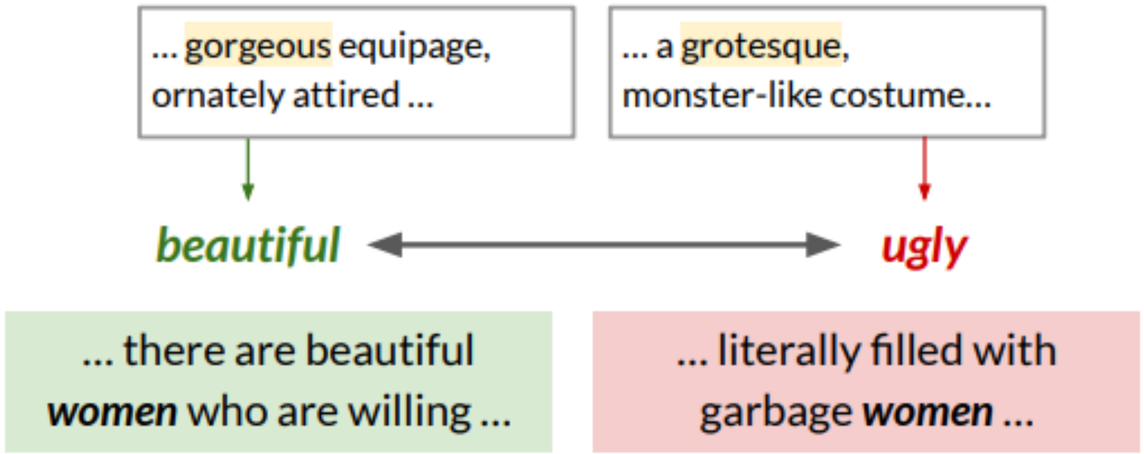


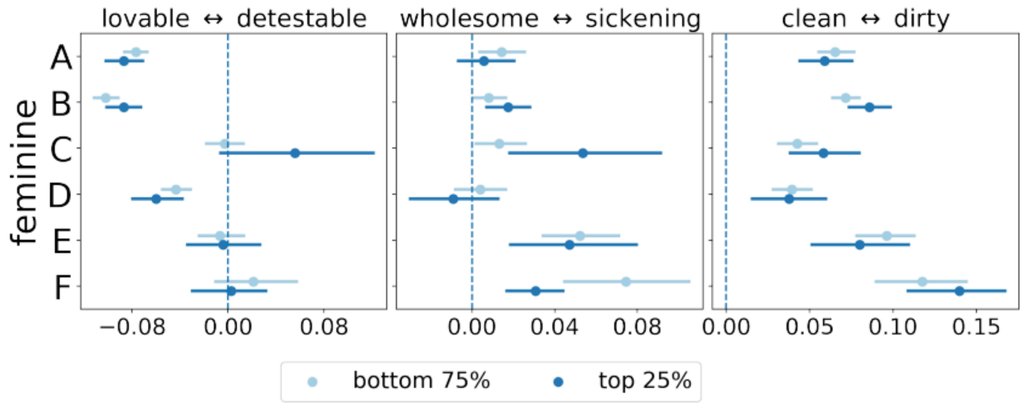
(a) Hate → hate

(b) Counterspeech → hate

He, Bing, et al. "Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis." *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2021.

Contextualized SemAxes





Li Lucy, Divya Tadimeti, and David Bamman. 2022. Discovering Differences in the Representation of People using Contextualized Semantic Axes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3477–3494, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

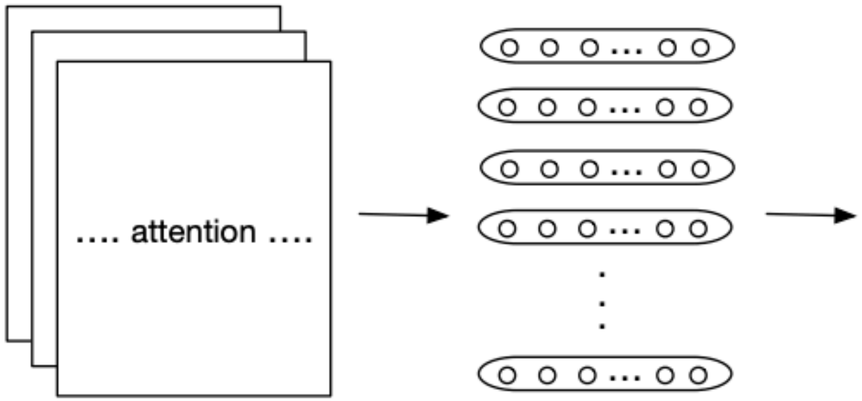
Geometry of contextualized Representations

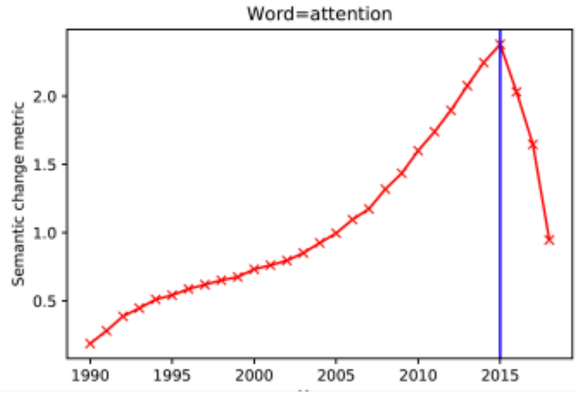
Region	North America	Europe	Middle East	Asia	South America	Oceania	Central America	Africa
BERT-Base	100%	92%	92%	91%	89%	87%	85%	85%
BERT-Base (Artificial Dataset)	100%	89%	92%	89%	88%	88%	87%	87%
BERT-Multilingual	100%	89%	88%	88%	91%	81%	83%	83%

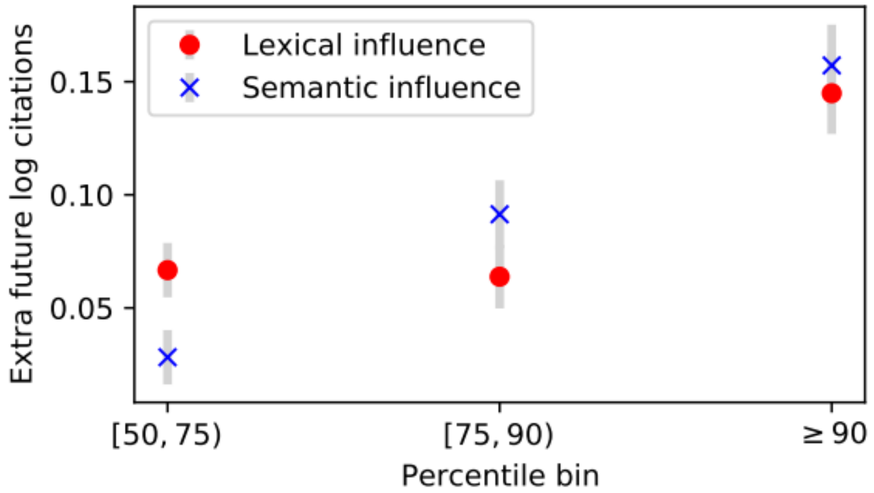
Table 1: % of the average radius of bounding balls relative to the average of radius of bounding balls of North American countries names. Central America also includes countries in the Caribbean.

Zhou et. al. 2021: https://arxiv.org/abs/2104.08465

Change over time







Soni, Sandeep, David Bamman, and Jacob Eisenstein. "Predicting Long-Term Citations from Short-Term Linguistic Influence." *Findings of the Association for Computational Linguistics: EMNLP 2022*. 2022.

In-class

- BERT Token Embeddings