Contextual Embeddings

Sandeep Soni

03/19/2024

Recap

- Language modeling: Predict next word based on some previous context i.e $P(x_i | x_1, x_2, \ldots, x_{i-1})$
- N-gram Los: Predict next word based on only the n-1 previous words
  - e.g., trigram (n=3) -> $P(x_i | x_{i-2}, x_{i-1})$
- Autoregressive LMs: Predict the next word by calculating the similarity between word embeddings and context embeddings
  - $P(x_i | x_1, x_2, \ldots, x_{i-1}) = \text{softmax}(\beta_{x_i} \cdot h_{i-1})$, where $h_{i-1} = \text{RNN}(x_{i-1}, h_{i-2})$

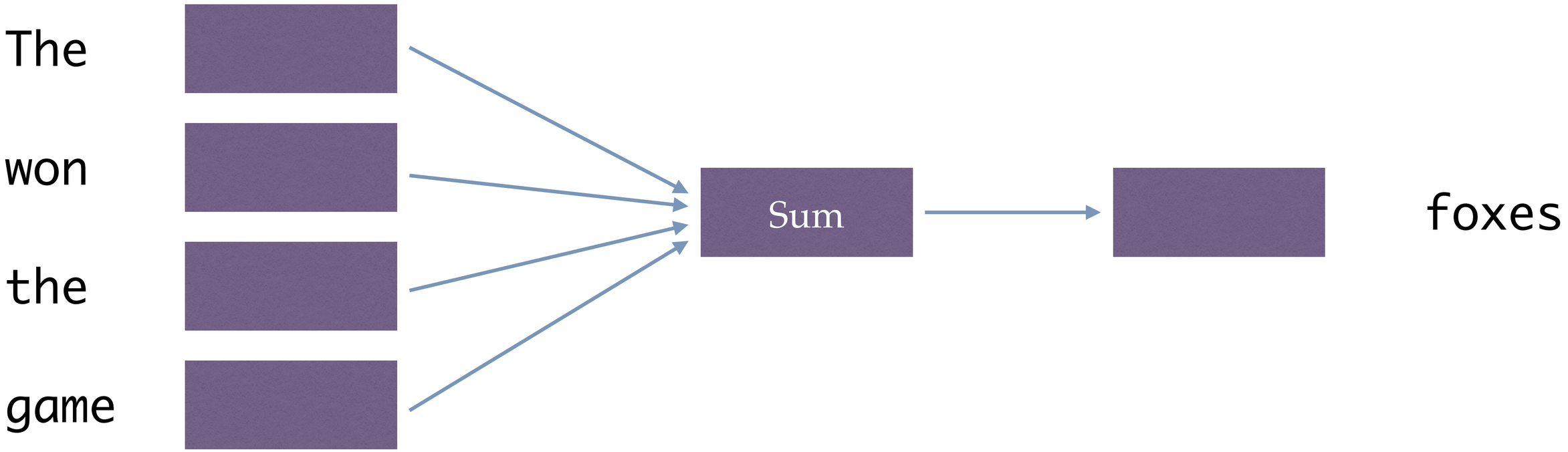
Masked Language modeling

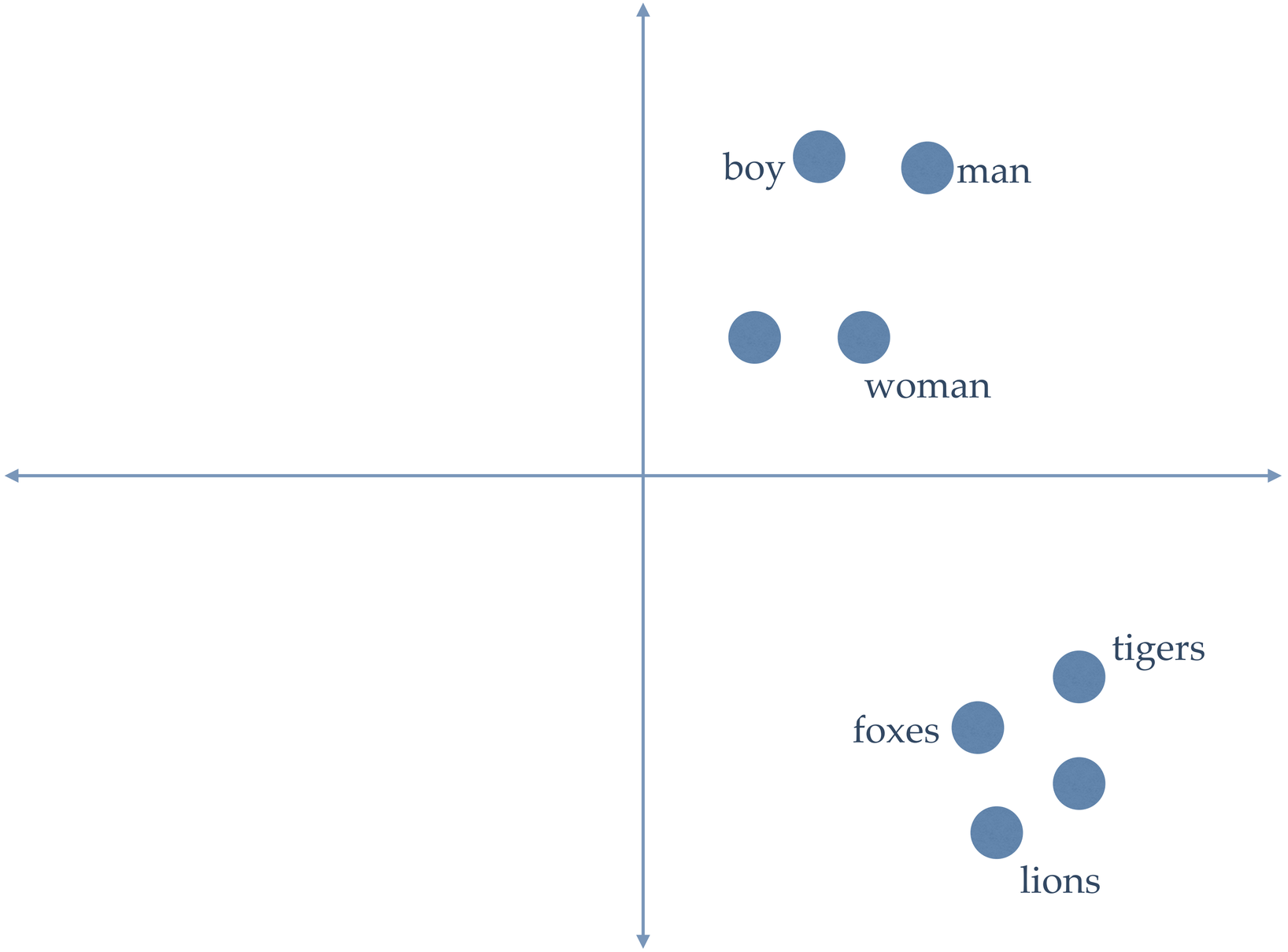
The _____ won the game

Fill in the blank by using the surrounding context

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

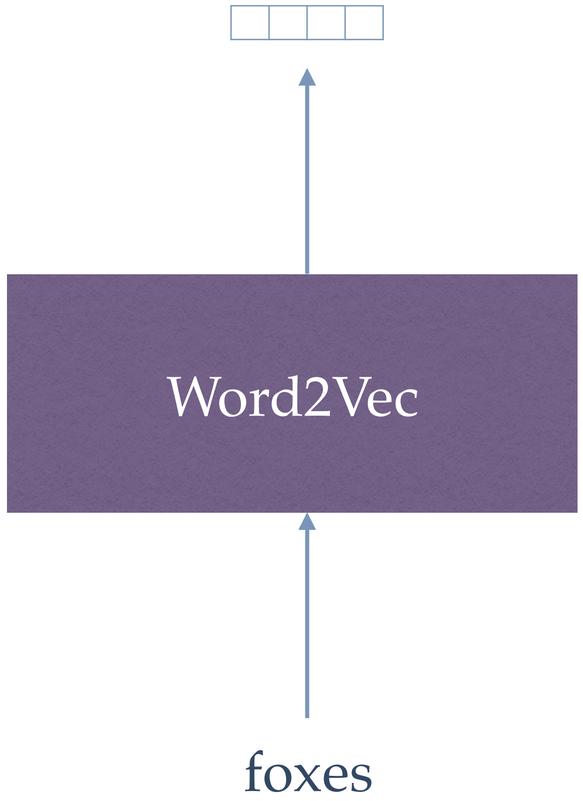
Word2Vec: CBoW





Semantic similarity transfers into geometric similarity

Word2vec abstraction



Input: word

Output: embedding

What is a limitation of word2vec?

- The foxes ran at us

- The foxes won the soccer game

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

Types & Tokens

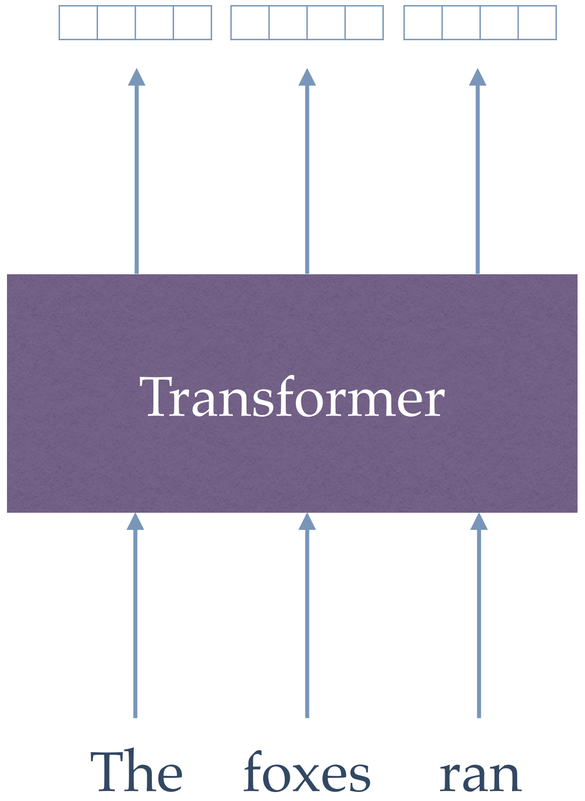
- Type: foxes
- Tokens:
  - The foxes ran at us

1.2	-0.1	0.7	-0.5

  - The foxes won the soccer game

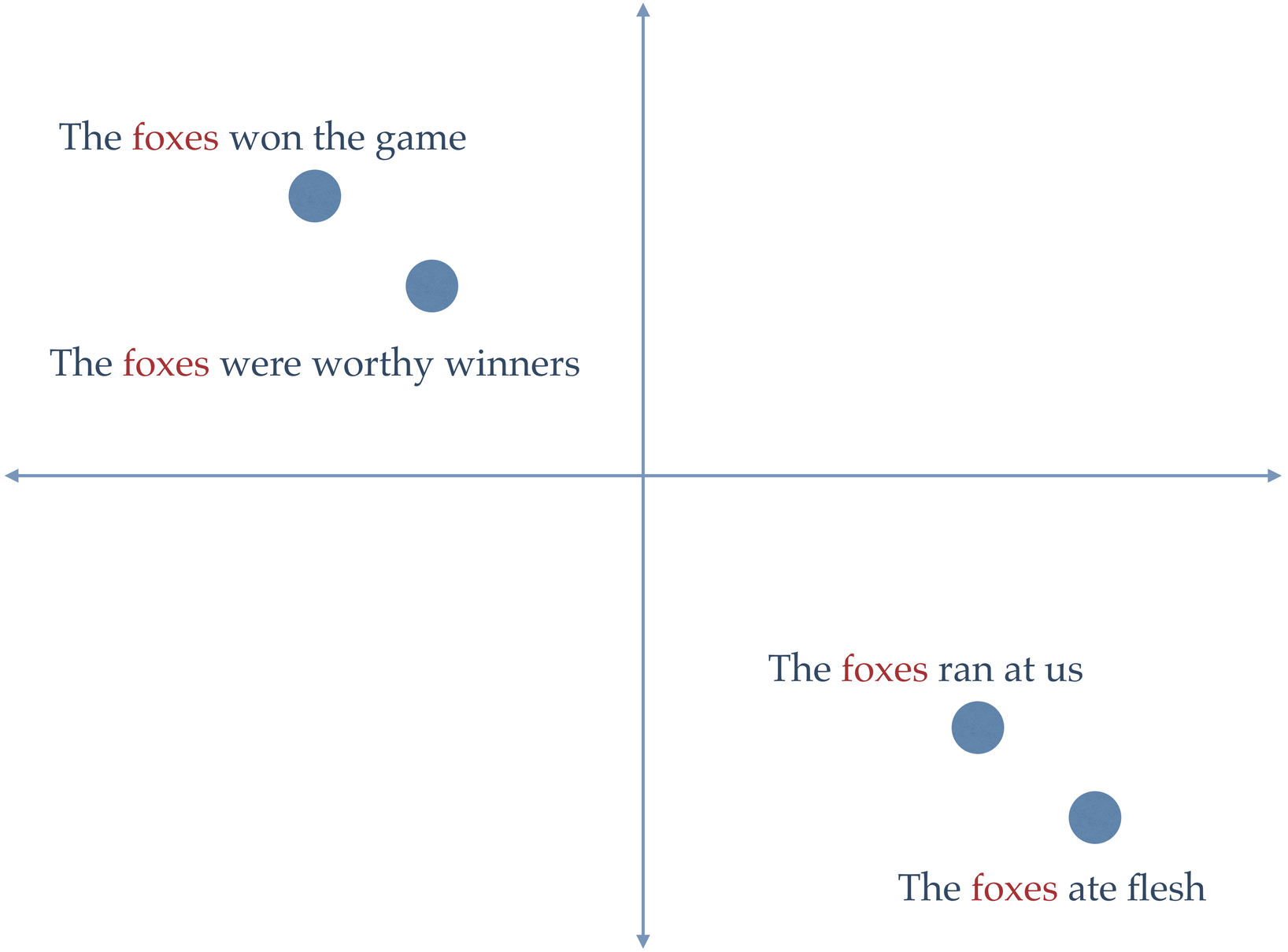
-0.4	0.6	0.8	0.4

Transformers: 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

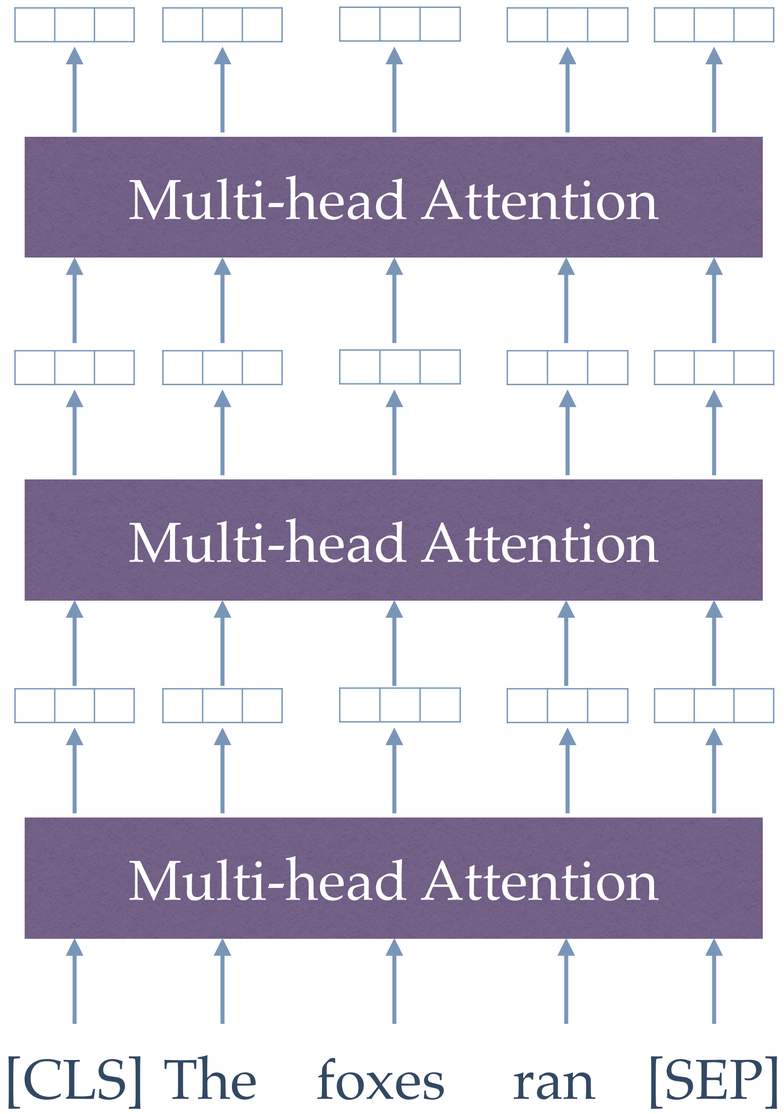
We further want these embeddings to be learned from training data and which can be updated for any task

BERT

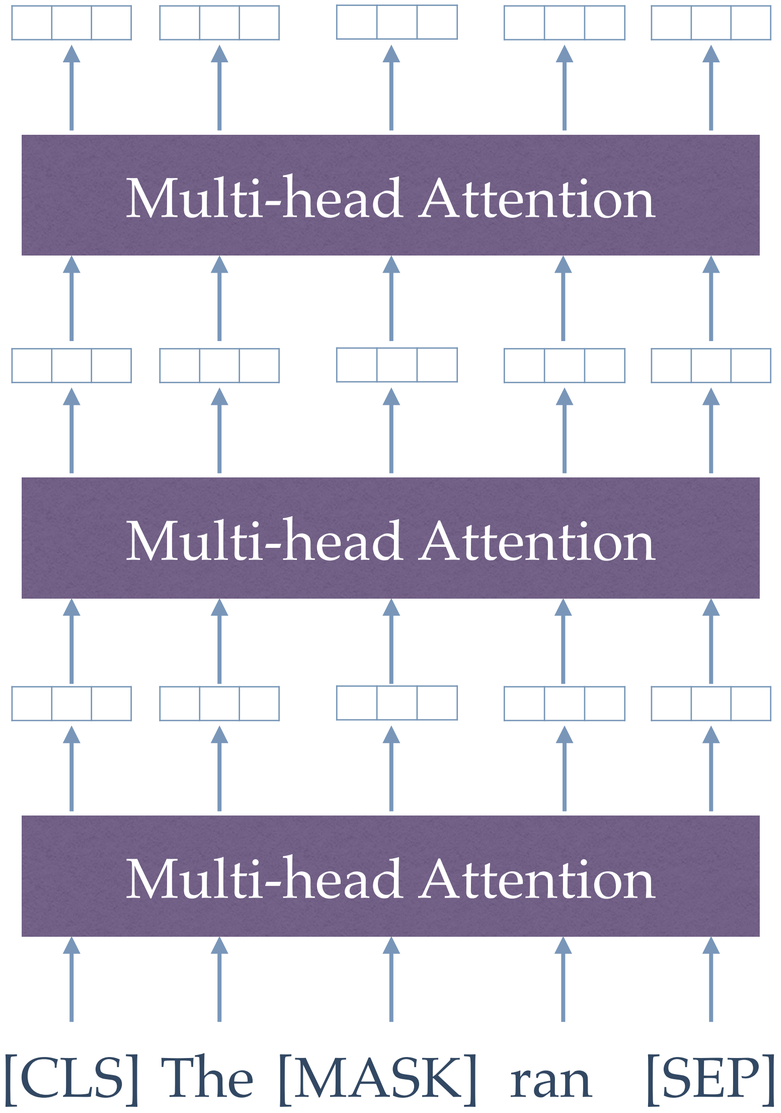
- Transformer based model that predicts masked word based on bidirectional context and next sentence prediction
- Multiple transformer blocks that take sequence of input vectors and give sequence of output vectors

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

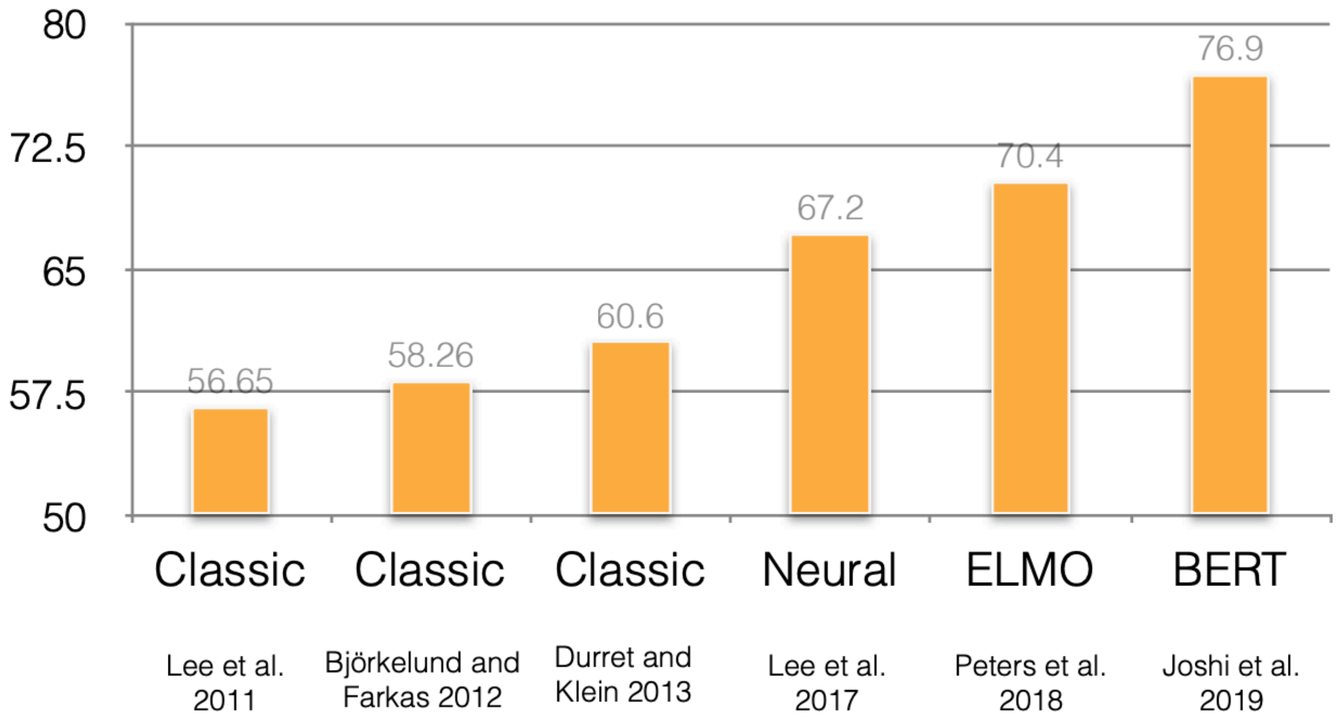
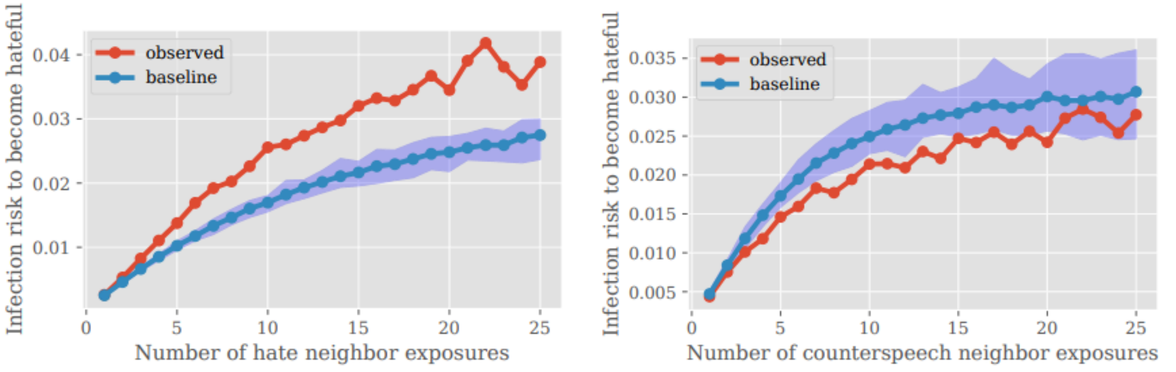


Figure taken from David Bamman's slide

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.

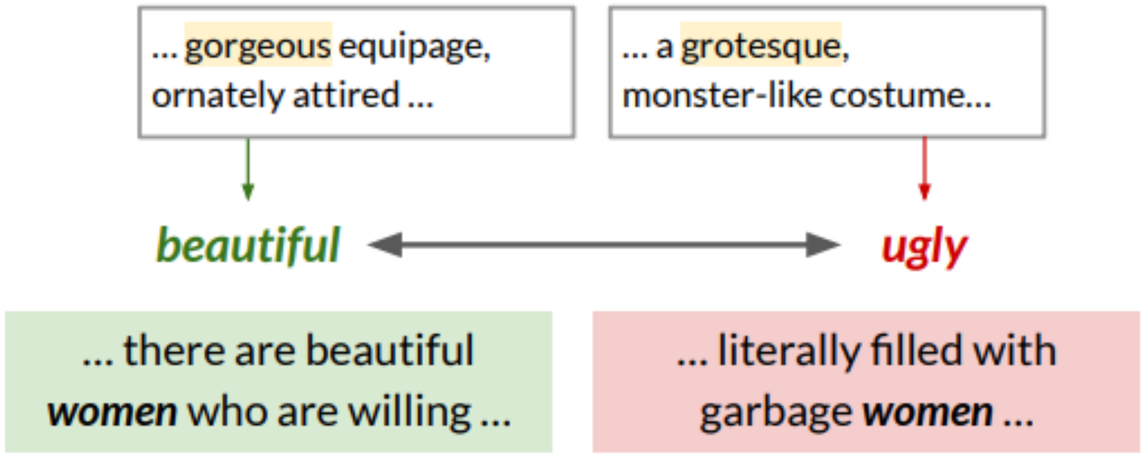


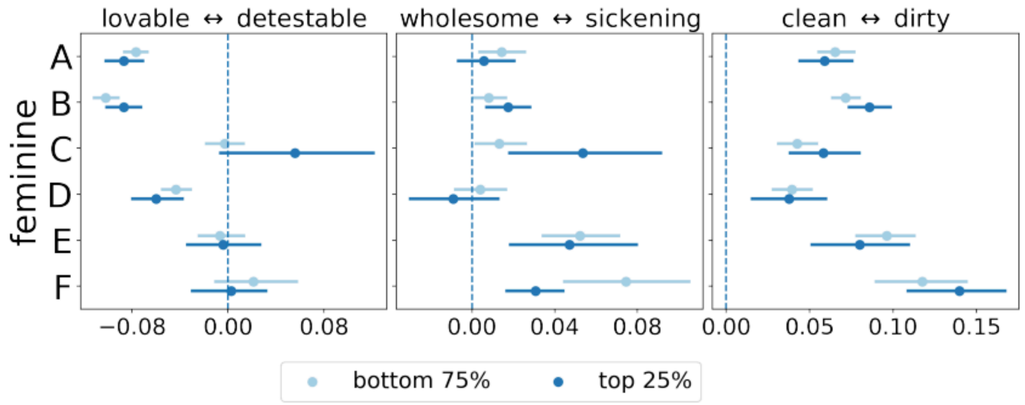
(a) Hate → hate

(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.

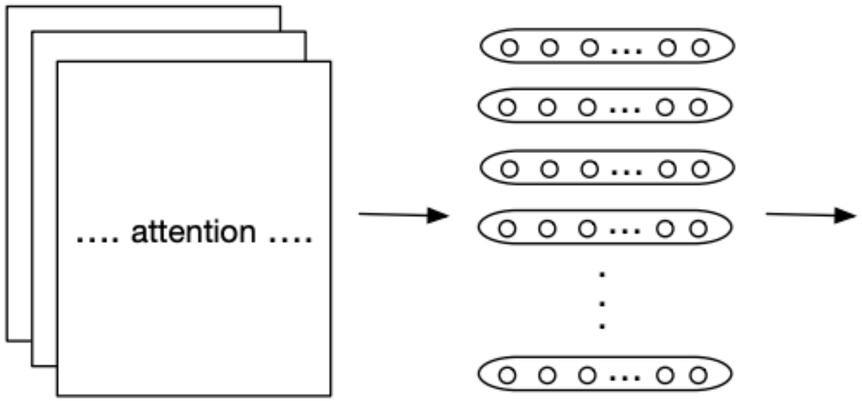
Geoemtry of contextualized Representations

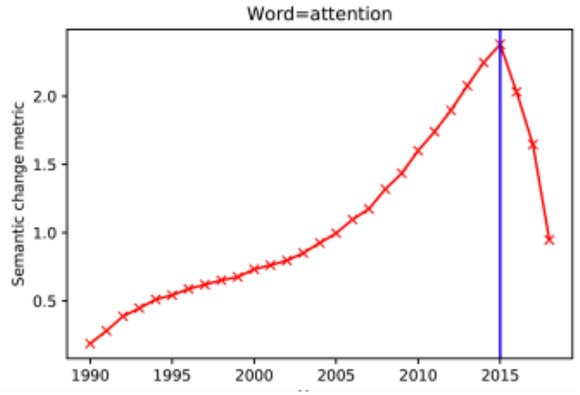
Region	North America	Europe	Middle East	Asia	South America	Oceania	Central America	Africa
BERT-Base	100%	92%	92%	91%	89%	87%	85%	85%
BERT-Base (Artificial Dataset)	100%	89%	92%	89%	88%	88%	87%	87%
BERT-Multilingual	100%	89%	88%	88%	91%	81%	83%	83%

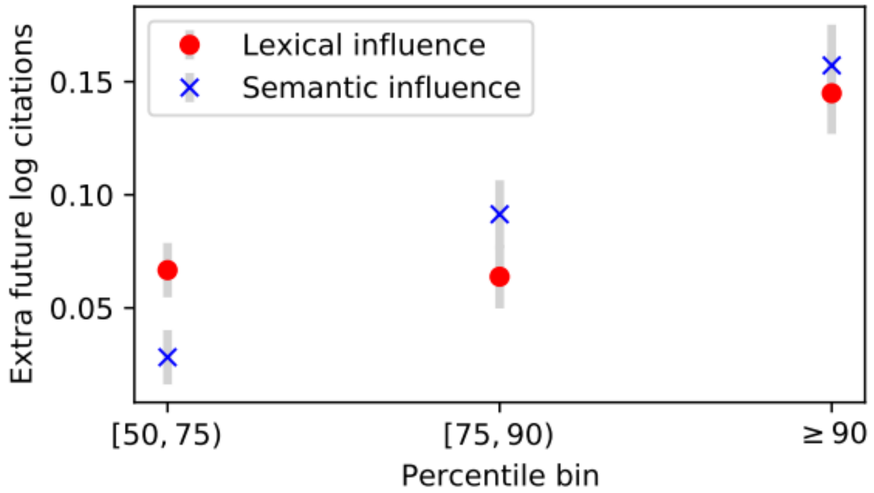
Table 1: % of the average radius of bounding balls relative to the average of radius of bounding balls of North American countries names. Central America also includes countries in the Caribbean.

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