University of BRISTOL

7.4 Properties of Word Embeddings

Edwin Simpson

Department of Computer Science,

University of Bristol, UK.

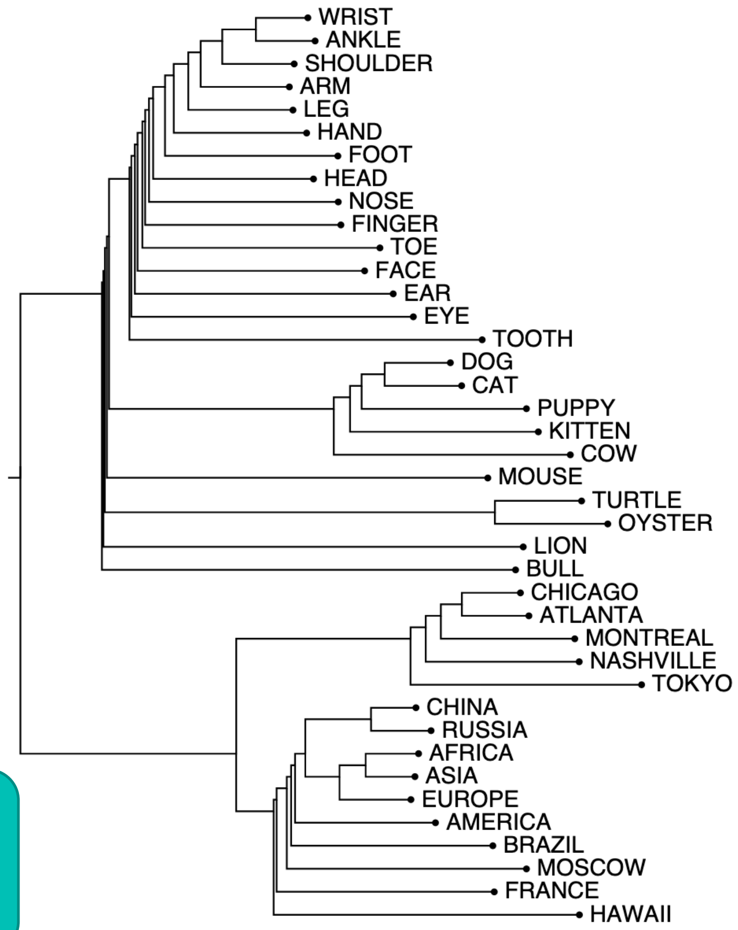
Context Window Size

- Typically 3-20 words
- Short:
  - Similar embeddings → related syntactic roles, e.g., noun/verb/etc.
  - *Hogwarts, Sunnydale*
- Long:
  - Similar embeddings → similar topics
  - *Hogwarts, Dumbledore*

Levy, O. and Goldberg, Y. (2014a). Dependency-based word embeddings. ACL.
Sections 6.9-6.13, Speech & Language Processing, 3rd edition draft, Jurafsky & Martin (2020).

Visualising Embeddings

- Run **principal component analysis (PCA)** to project the embeddings into a few dimensions, then plot pairs of dimensions.
- Hierarchical clustering: produce a tree of relations between terms →
- Visualisation can be useful to discover semantic properties.



Rohde, D. L., Gonnerman, L. M., & Plaut, D. C. (2006). An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*,

Semantic Relations

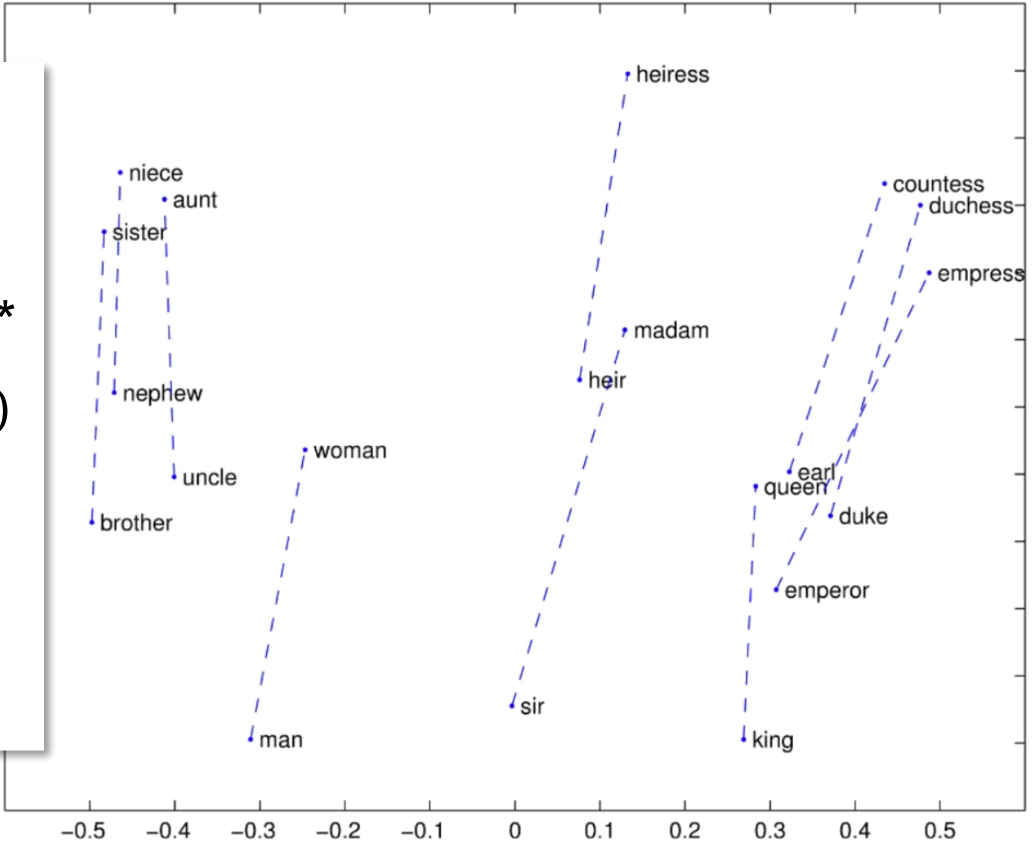
Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. EMNLP 2014

- Offsets between embeddings capture semantic relations
- Gender →
- Analogy: a is to b as a* is to b*

$$\widehat{b^*} = \text{argmax}_{b^*} sim(b^*, b - a + a^*)$$

$$king - man + woman \approx queen$$

- Often works only if we exclude variants of the input words.

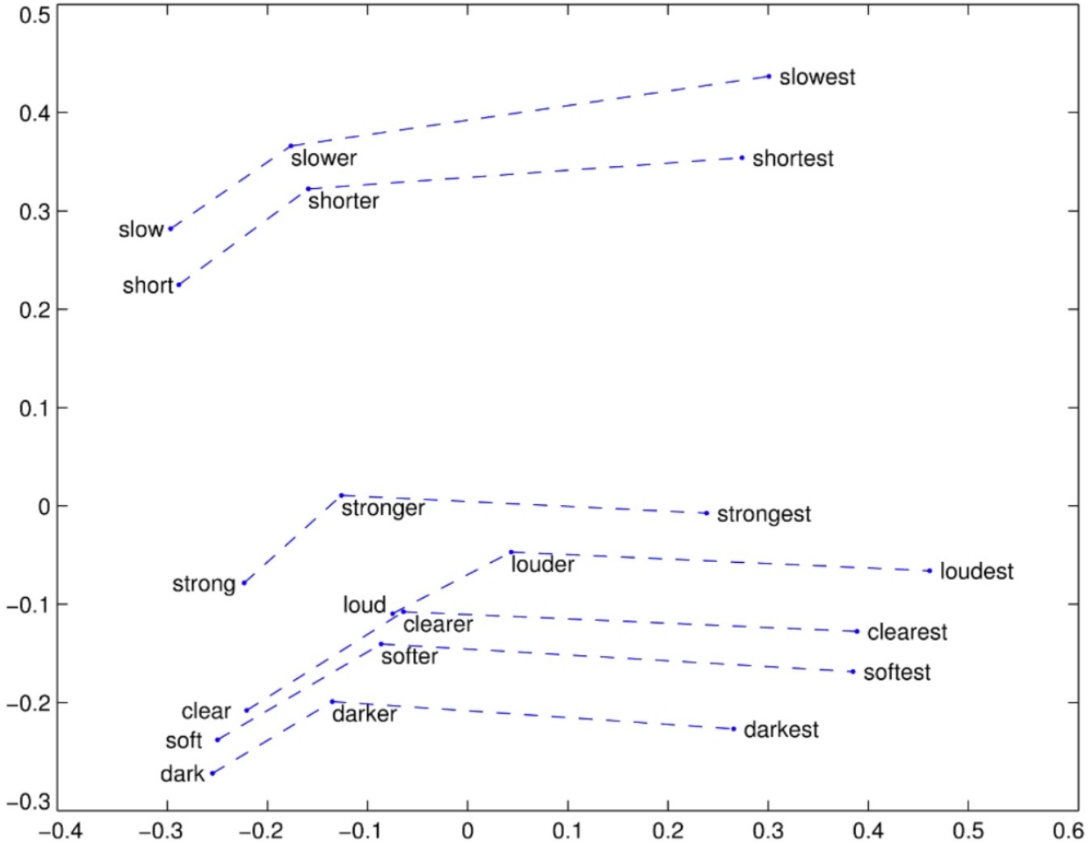


Semantic Relations

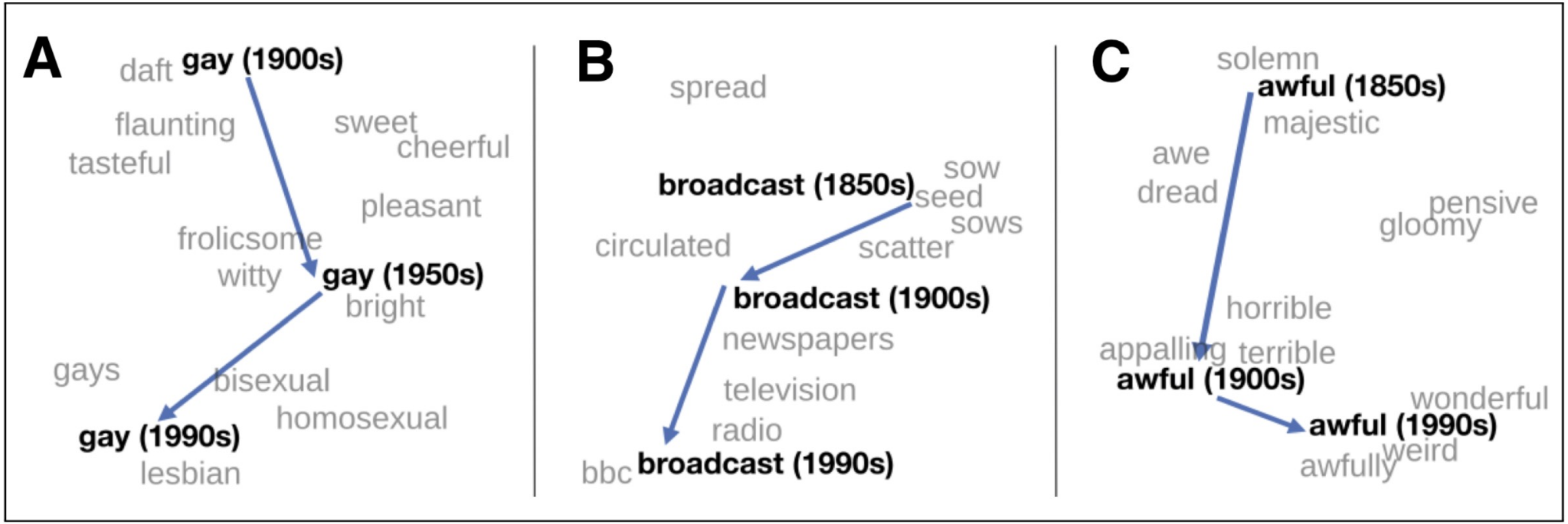
- Offsets between embeddings capture semantic relations
- Comparatives and superlatives →

$$shorter - short + slow \approx slower$$

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. EMNLP 2014



Tracking Embeddings Over Time



Unwanted Bias in Embeddings

- Embeddings learn **biases** from the data they are trained on.
- E.g., unwanted associations such as jobs with particular groups of people
- Modelling assumptions can also introduce bias without the developers being aware of it:
  - E.g., embeddings have a limited capacity to encode information, so compressing models to a smaller size saves memory/computation cost
  - However, it also causes them to forget information about rarer terms and contexts, which adversely affects underrepresented groups.

Hooker, S., Moorosi, N., Clark, G., Bengio, S., & Denton, E. (2020). Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*.

Unwanted Bias in Embeddings

- E.g., associating certain occupations with gender
- Use the analogy method on word2vec embeddings trained on Google News corpus:

$$\overrightarrow{computer\ programmer} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{homemaker}$$

- Analogy generator:
  - Inputs: two words, e.g., ‘man’, ‘woman’;
  - Outputs: two **different** words with a similar offset

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *NeurIPS*.

Implications of Bias

- Biased embeddings cause unfairness and errors in downstream tasks:
  - **Job candidate search**: automatic CV rating algorithm may down-weight women's names for computer programmer roles;
  - **Sentiment analysis**: stronger association of African-American names with unpleasant words than European-American names → algorithm predicts more negative sentiment toward African-Americans;
  - **Coreference resolution:** in a sentence, does '*the programmer*' refer to the man or the woman that was mentioned earlier in the sentence?
  - **Translation:** gender-neutral terms are often translated to male gender.

Debiasing Embeddings

An open problem;

1. Use a set of defining words to identify a subspace of embeddings, *B* (e.g. a direction), corresponding to a bias.
2. Neutralise: to debias an embedding, *w*, of a gender-neutral word (e.g., ‘nurse’), set its value in the subspace to zero.
3. Define sets of words relating to different genders, e.g., *{guy,gal}.*
4. Equalise: reposition the neutralised *w* so that it is equidistant from words in the defining set outside *B*.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *NeurIPS*.

Evaluation

- Similarity:
  - Use datasets with human-assigned ratings for pairs of words.
  - Some judge words in isolation, others judge in context to account for variations in meaning.
  - Compute cosine similarity scores for the embeddings, then compute their correlation with human judgements.
- Downstream tasks:
  - Estimate performance on a downstream task like sentiment analysis or information extraction.

Summary

- Context window size affects the semantics of embeddings.
- Visualisation using hierarchical clustering or PCA helps to identify semantic properties of embeddings and changing meaning.
- Properties include gender, comparatives & superlatives.
- Unwanted biases are learned from data, which we can try to remove using debiasing methods with limited success.
- Evaluation is either intrinsic (word pair similarity) or extrinsic (downstream task performance).