Model Choices Influence Attributive Word Associations

A Semi-supervised Analysis of Static Word Embeddings

Geetanjali Bihani, Julia Taylor Rayz

Purdue University

Purdue University | Polytechnic Institute

AKRaNLU

Paper Snapshot

Goal

- Analyze how embedding training model choices impact attributive word associations

Motivation

- Word associations not explicitly encoded in word vector spaces created by off-the-shelf shallow NNs (word2vec, GloVe, fastText, etc.)
- Variation in word associations based on embedding training procedure

Approach

- Semi-supervised cluster analysis on annotated proper nouns and adjectives based on word embedding features
- Reveals changes in developed attributive word associations and the embedding space

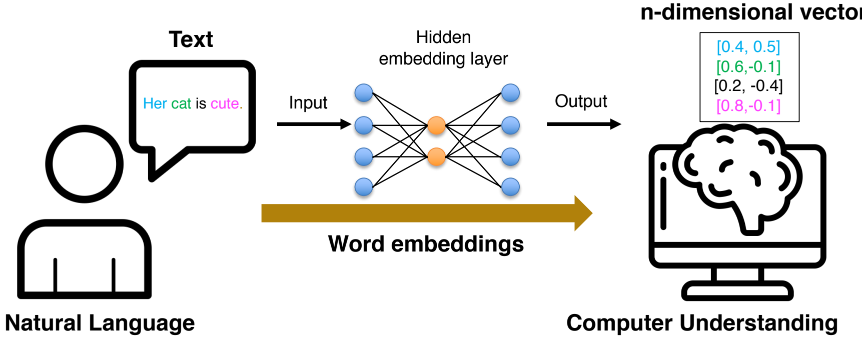
Findings

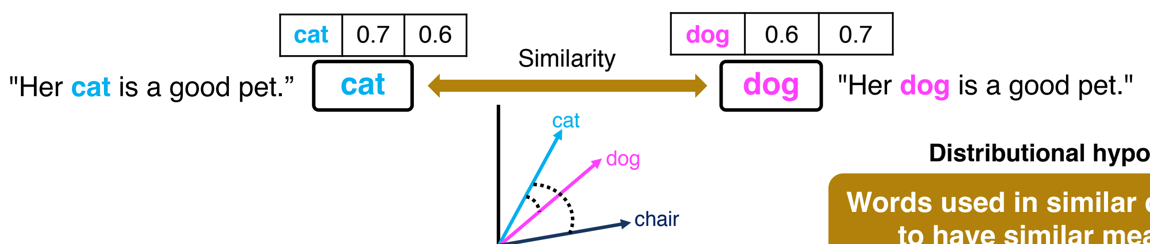
- Choice of context learning flavor (CBOW vs skip-gram) impacts distinguishability and sensitivity of word embeddings towards training corpora
- Significant inter-model disparity and intra-model similarity in word associations, when trained over same corpora

Introduction

- **Word Embeddings: Map words as n–dimensional vectors**

  Each word stored as a point in space, as a vector of a fixed number of dimensions





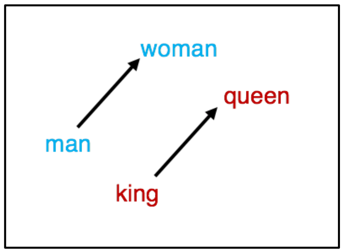
Distributional hypothesis

Words used in similar contexts tend to have similar meanings [1]

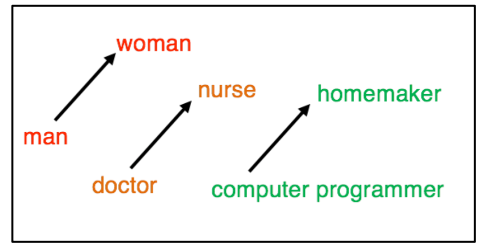
[1] Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162.

Word Associations

- **Word embedding models encode…**



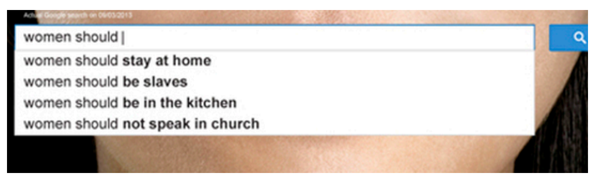
Semantic and syntactic regularities in language [2]



Undesirable word associations [3], [4]

- **Raises concerns regarding the validity of application of these models in the real-word**

Crime recidivism prediction



Recommendation engines

Microsoft shuts down AI chatbot after it turned into a Nazi

AI chatbots

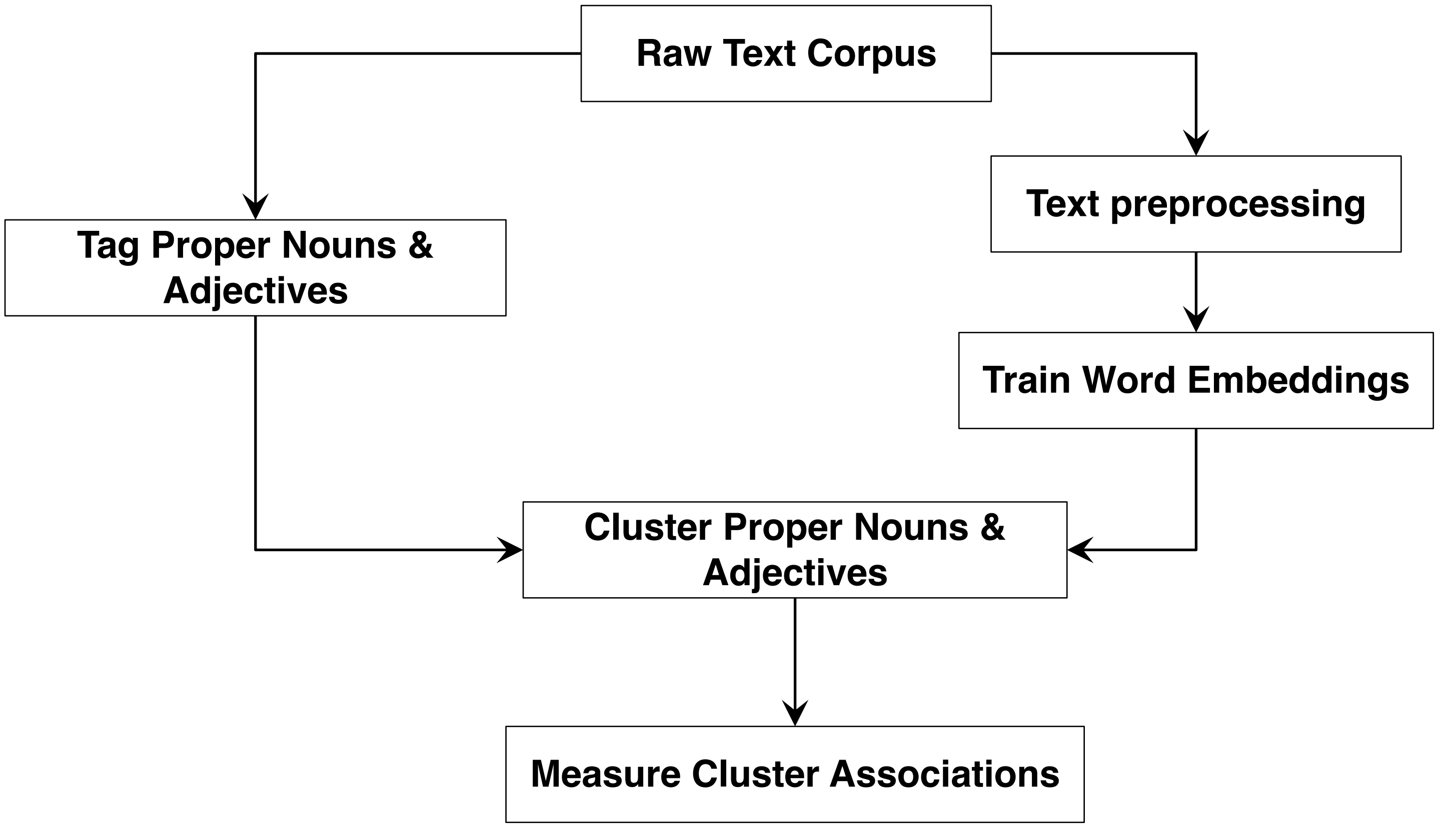
[2] Mikolov, T., Yih, W. T., & Zweig, G. (2013, June). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746-751).
[3] Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.
[4] 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. In Advances in neural information processing systems (pp. 4349-4357).

Word Associations

- Word associations in various word embedding architectures trained on different text corpora not comparable [5]
- Search for explicitly defined ‘biased’ word associations [6]
    - Ignores other biases (gender, religious, racial, etc.)
    - Introduce researcher’s cultural biases regarding certain concepts
- Crucial to assess variations in word associations across different embedding model choices
    - Model architecture
    - Training corpora
    - Context learning process
- **How do model architecture and corpus choices influence word associations?**

[5] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018, June). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (pp. 15-20).
[6] Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.

Methodology



Methodology

- **Data:** Corpus of Historical American English [7]

  ~ 1.5 million unique words in corpus

- **Neutral and Attribute words**

  Neutral: Proper Nouns I **Attribute**: Adjectives

  Word labeling done using the Stanza NLP POS tagger [8]

- **Text Preprocessing**

  Remove special characters, numeric characters and special spaces, Lowercase tokens

Decade	Total Number Of Words	Unique words
1810s	1,181,022	10,110
1820s	6,927,005	28,925
1830s	13,773,987	45,154
1840s	16,046,854	49,311
1850s	16,493,826	48,866
1860s	17,125,102	58,080
1870s	18,610,160	53,991
1880s	20,872,855	59,489
1890s	21,183,383	65,742
1900s	22,541,232	73,628
1910s	22,655,252	67,200
1920s	25,632,411	84,259
1930s	24,413,247	95,032
1940s	24,144,478	95,040
1950s	24,398,180	101,078
1960s	23,927,982	97,827
1970s	23,769,305	102,356
1980s	25,178,952	109,878
1990s	27,877,340	116,459
2000s	29,479,451	123,323
Totals	406,232,024	1,485,748

[7] Davies, M. (2015). Corpus of Historical American English (COHA) [linguistic corpora]. Retrieved from: https://doi.org/10.7910/DVN/8SRSYK

[8] Peng Qi, Timothy Dozat, Yuhao Zhang and Christopher D. Manning. 2018. Universal Dependency Parsing from Scratch In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp. 160-170.

Word Embedding Models

Word Embedding	Word context	Co-occurrence matrix	Character n-grams
Word2vec (2013) (CBOW & Skip-gram) [9]	✅	❌	❌
GloVe (2014) [10]	✅	✅	❌
Fasttext (2016) (CBOW & Skip-gram) [11]	✅	❌	✅

[9] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*

[10] Pennington, J., Socher, R., & Manning, C. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

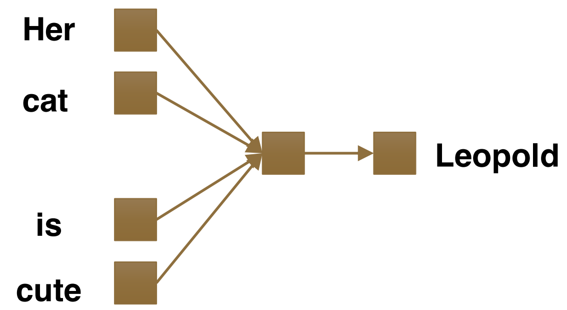
[11] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146.

Model Hyperparameters

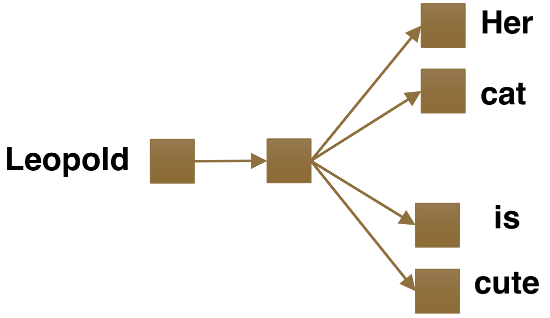
Parameter	Word2vec (CBOW)	Word2vec (Skip-gram)	GloVe	fastText (CBOW)	fastText (Skip-gram)
window	5	5	5	5	5
model	sg = 0 [cbow]	sg = 1 [skip-gram]	-	“cbow”	“skipgram”
alpha	0.025	0.025	0.025	0.025	0.025
max_vocab_size	None	None			
epochs	1	1	1	1	1

CBOW vs Skip-gram

- **CBOW**: The network tries to **predict which word** is most likely, **given its neighboring words (context).**



- **Skip-gram**: The network tries to **predict neighboring words (context)** which are most likely, **given the current word**.

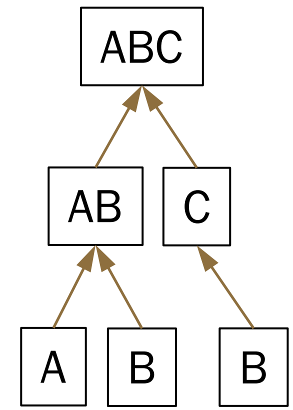


Noun-Adjective Clustering

- **Agglomerative Hierarchical Clustering**
- **Distance Measure:** Cosine distance $D_c(A,B)$
- **Linkage criteria: Ward linkage**

  Accounts for merging cost of combining a pair of clusters

  Uncovers non-round and non-uniform clusters



- **Merging cost:**

$$\Delta(A,B) = \sum_{i \in A \cup B} |\overrightarrow{x_i} - \overrightarrow{m_{A \cup B}}|^2 - \sum_{i \in A} |\overrightarrow{x_i} - \overrightarrow{m_A}|^2 - \sum_{i \in B} |\overrightarrow{x_i} - \overrightarrow{m_B}|^2$$

$$\Delta(A,B) = \frac{n_A n_B}{n_A + n_B} |\overrightarrow{m_A} - \overrightarrow{m_B}|^2$$

Where $\overrightarrow{m_j}$ is the center of cluster j, and n_j is the number of points in it.

Δ is called the merging cost of combining cluster A and B

Optimal No. of word clusters

- Computational heuristic methods don't identify a clear preference of number of clusters
- Utilized theories informing distinctions between adjectives

Adjectives have **semantic orientations** and **gradability** attached [12]

Semantic orientations

positive, negative and neutral

Gradability

Comparative constructs

Root morphemes of adjective words can be traced to emotions [13]

fearful → fear → fear

amazing → amaze → amazement

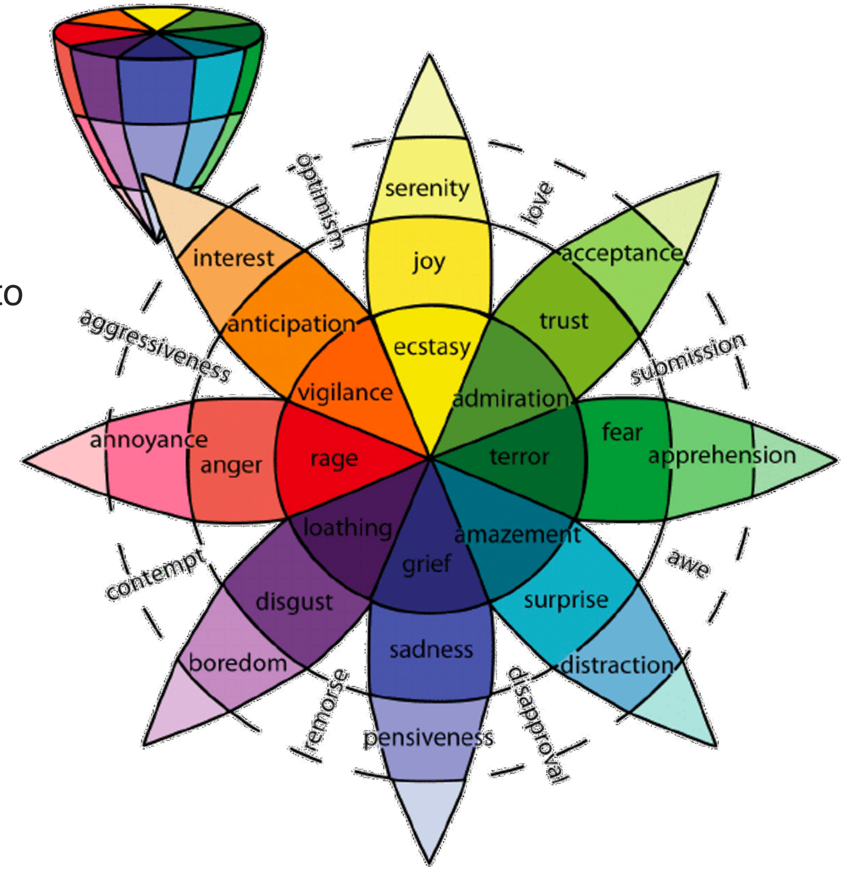
[12] Kim, E., & Klinger, R. (2018). A survey on sentiment and emotion analysis for computational literary studies. arXiv preprint arXiv:1808.03137.
[13] Johnson-Laird, P. N., & Oatley, K. (1989). The language of emotions: An analysis of a semantic field. Cognition and emotion, 3(2), 81-123.

Optimal No. of word clusters

- Extended the semantic orientations of adjectives to emotion space

  Plutchik's wheel of emotions: Framework for distinguishing emotions [14]

  Emotions represented by most adjectives traced back to the **8 prototype emotion themes**



[14] Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In Theories of emotion (pp. 3-33): Elsevier.

Measuring Word Associations

- **Inter-cluster**

Dunn's Index: Used to assess cluster validity [15]

$$Du(K) = \min_{i=1,\ldots,K}\left(\min_{j=i+1,\ldots,K}\left(\frac{D(C_i, C_j)}{\max_{l=1,\ldots,K} diam\,(C_l)}\right)\right)$$

$$D(C_i, C_j) = \min_{\boldsymbol{x}\in C_i, \boldsymbol{y}\in C_j} D\,(\boldsymbol{x}, \boldsymbol{y})$$

$$diam(C_i) = \max_{\boldsymbol{xy}\in C_i} D\,(\boldsymbol{x}, \boldsymbol{y})$$

- **Intra-cluster**

Distribution of words within clusters: Proportion of words in each cluster

Jaccard Similarity: Clusters compared across each of the 5 WE model clusters

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

$$Average\ J(C_A, C_B)_{1810s-2000s} = \frac{J(C_A, C_B)_{1810s} + J(C_A, C_B)_{1820s} + \cdots + JJ(C_A, C_B)_{2000s}}{20}$$

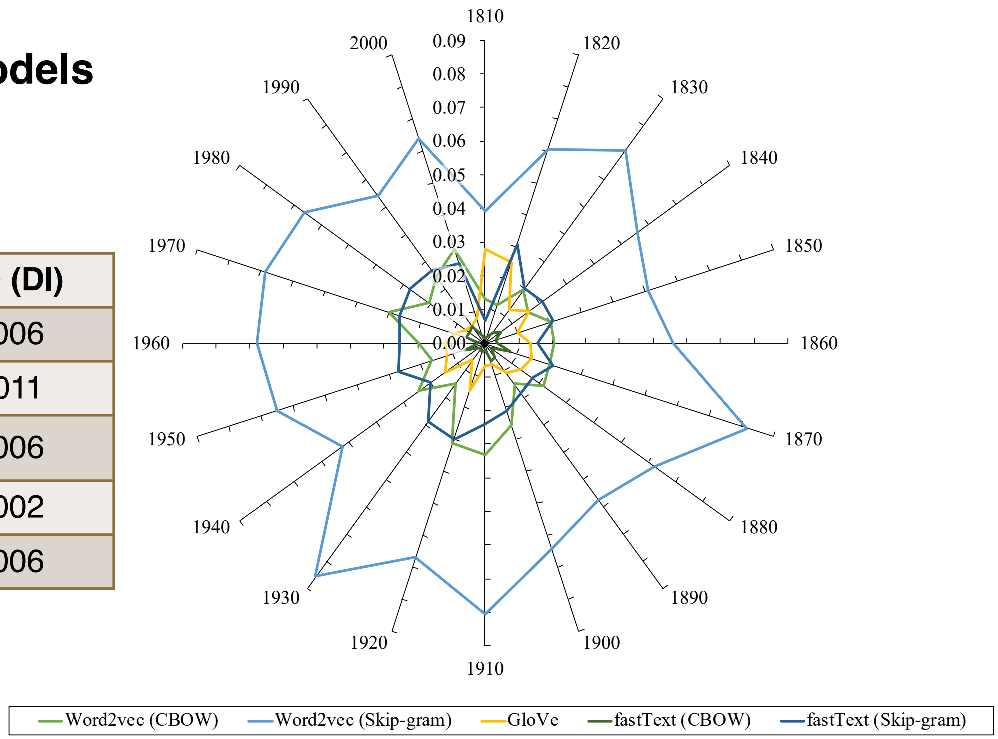
[15] Xu, R. (2015). Clustering: Piscataway, New Jersey : IEEE Press.

Results

- **Variation in Dunn's Index across models**

  word2vec > fastText > GloVe

Model	M^a (DI)	SD^a (DI)
word2vec (CBOW)	0.021	0.006
word2vec (skip-gram)	0.063	0.011
GloVe	0.012	0.006
fastText (CBOW)	0.004	0.002
fastText (skip-gram)	0.023	0.006



- **Stronger distinctions between words captured for skip-gram as compared to CBOW**
- **Sensitivity towards lexical regularities higher in skip-gram as compared to CBOW**

Results

- **Variation in Dunn's Index across corpora**

	Word2vec (CBOW)	Word2vec (Skip-gram)	GloVe	fastText (CBOW)	fastText (Skip-gram)
1810	0.013	0.039	0.028	0.001	0.007
1820	0.012	0.061	0.025	0.002	0.031
1830	0.020	0.071	0.012	0.004	0.020
1840	0.016	0.056	0.016	0.006	0.021
1850	0.021	0.051	0.010	0.004	0.021
1860	0.021	0.056	0.014	0.004	0.016
1870	0.020	0.082	0.015	0.008	0.021
1880	0.022	0.062	0.013	0.002	0.017
1890	0.015	0.058	0.011	0.005	0.018
1900	0.026	0.064	0.007	0.005	0.021
1910	0.033	0.081	0.007	0.002	0.024
1920	0.031	0.067	0.015	0.003	0.030
1930	0.015	0.086	0.006	0.002	0.029
1940	0.024	0.052	0.015	0.002	0.020
1950	0.016	0.065	0.012	0.006	0.027
1960	0.020	0.068	0.012	0.002	0.025
1970	0.030	0.069	0.009	0.006	0.027
1980	0.021	0.066	0.007	0.006	0.028
1990	0.025	0.054	0.007	0.006	0.027
2000	0.029	0.064	0.008	0.002	0.025

- **Word associations become increasingly distinct as training corpus becomes larger**

Results

- **Distribution of words across clusters**

% of words/cluster	Word2vec (CBOW)	Word2vec (SG)	GloVe	fastText (CBOW)	fastText (SG)
Min.	0.12%	1.14%	0.03%	0.82%	1.77%
Mean	12.50%	12.50%	12.50%	12.61%	12.53%
Max.	82.34%	51.17%	79.81%	44.52%	34.14%

- **GloVe and word2vec (CBOW) embeddings encode minimal distinction between words**
- **Different word embedding models encode different vector spaces for the same training corpora**

Results

- **Jaccard Similarity**

Model	word2vec (CBOW)	word2vec (skip-gram)	GloVe	fastText (CBOW)	fastText (skip-gram)
word2vec (CBOW)	1	-	-	-	-
word2vec (skip-gram)	0.38	1	-	-	-
GloVe	0.44	0.36	1	-	-
fastText (CBOW)	0.34	0.46	0.35	1	-
fastText (skip-gram)	0.39	0.5	0.41	0.62	1.00

- **Difference in word context consideration drives differences in word associations**

- **word2vec embeddings more affected by changes in context-learning flavor, compared to fastText**

Conclusion

- Context learning architecture, type of embedding model and size of training corpora influence the embedding spaces generated

- **Context learning architecture** influences the sensitivity of word embeddings towards changes in training corpora

  skip-gram architecture captures more distinguishable word associations as compared to CBOW

- **Type of embedding model**

  word2vec encodes the most distinguishable word associations as compared to fastText and GloVe

- **Size of training corpora**

  Distinguishability of word associations increases when models trained on larger corpora*

  *except for GloVe, word association strength degraded possibly due to the limited number of training epochs