Discovering Stylistic Variations in Distributional Vector Space Models via Lexical Paraphrases

Xing Niu, and Marine Carpuat
CLIP Lab, University of Maryland

DEPARTMENT OF
COMPUTER SCIENCE

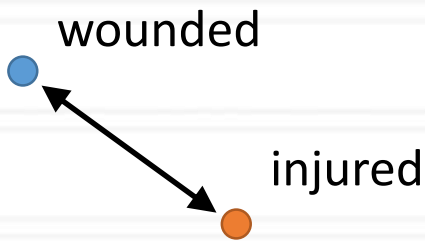
2017.09.08 @ EMNLP-Stylistic Variation

Vector Space Model & Word Similarity

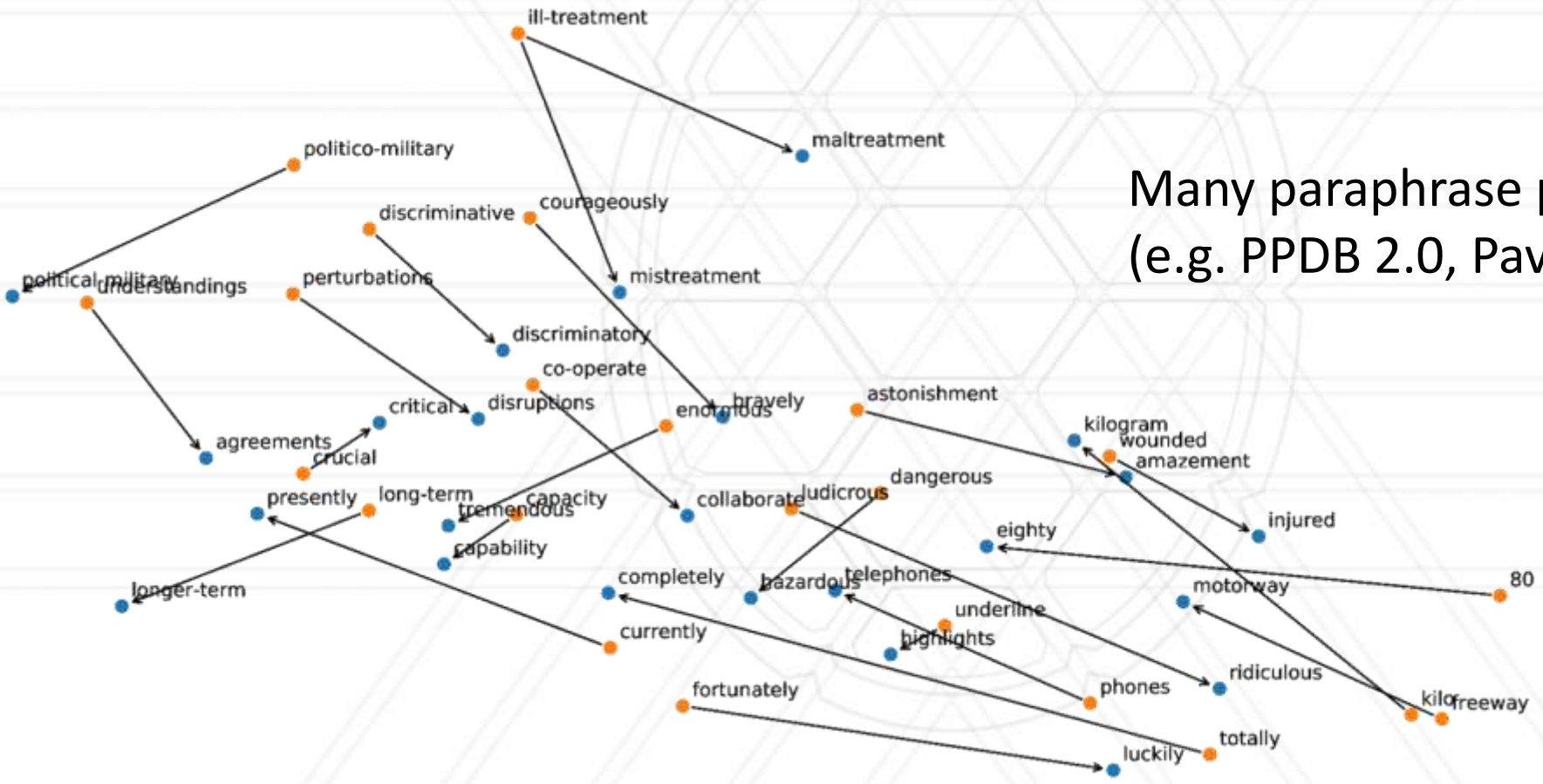
Vector Space Model & Word Similarity

synonyms/paraphrases
- same meaning
- non-zero distance vectors
- What difference remains?

- Style can vary along many dimensions.
- How can we robustly discover stylistic variations in vector space?

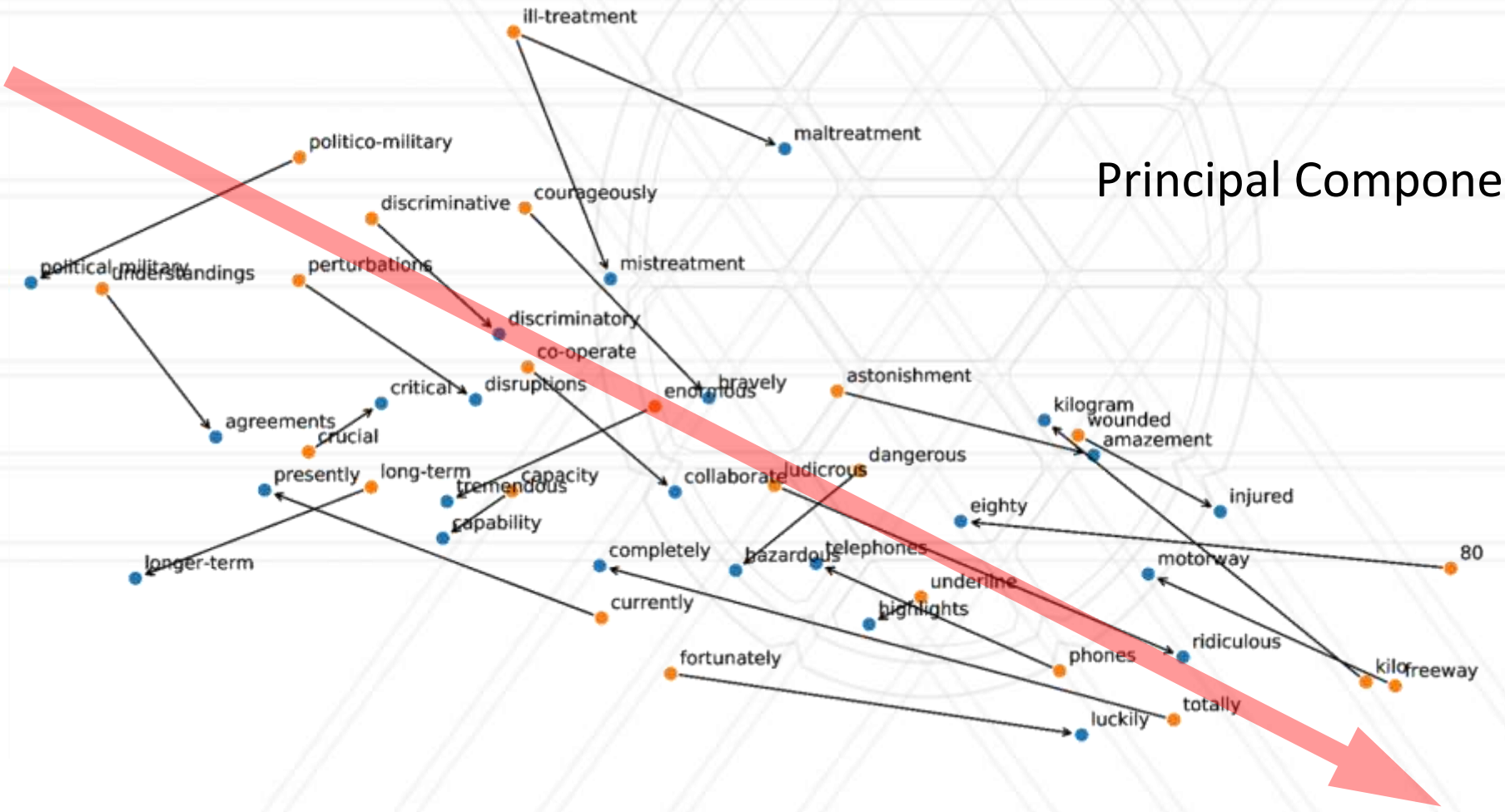


Discovering Stylistic Variations



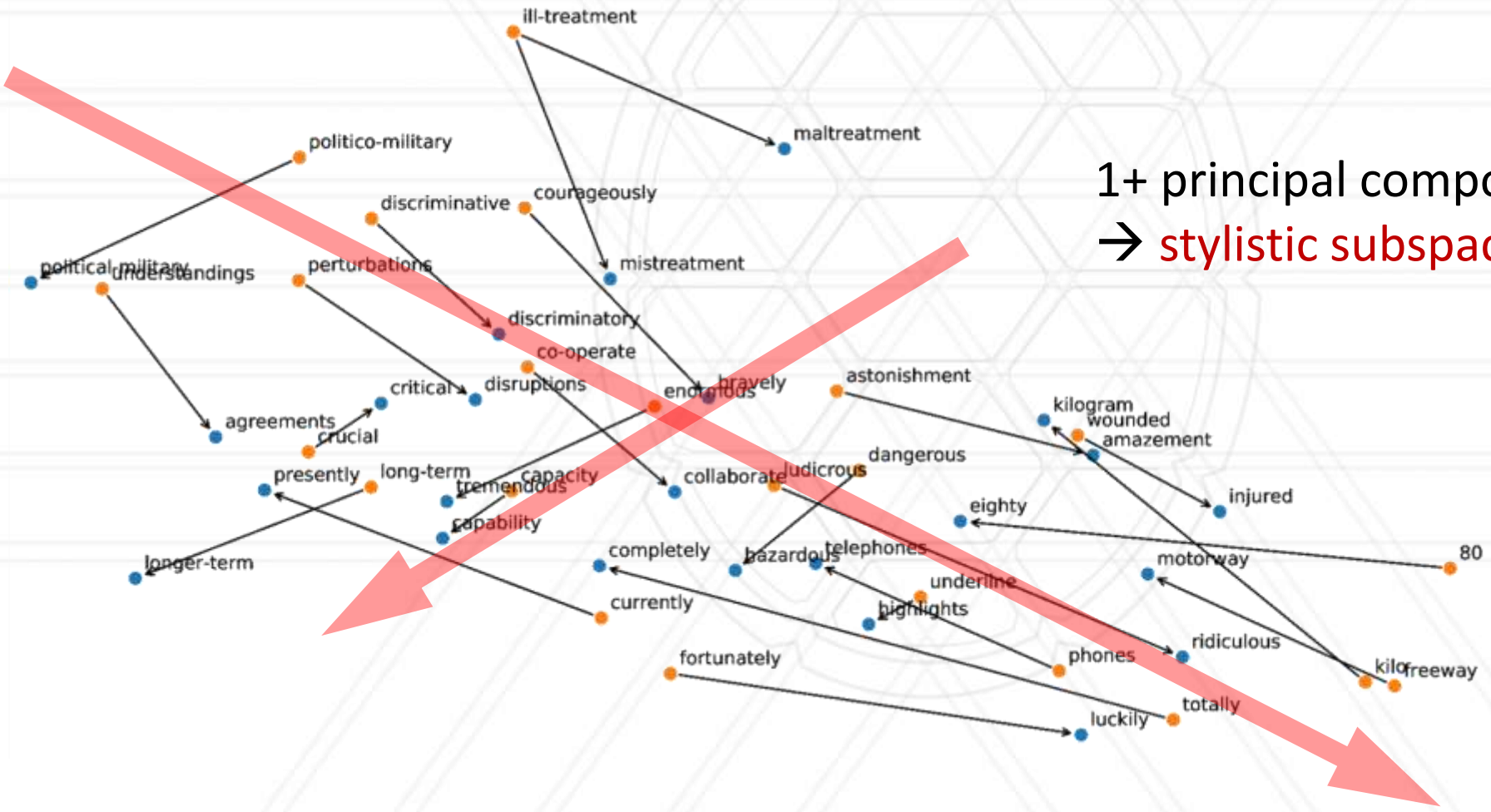
Many paraphrase pairs!
(e.g. PPDB 2.0, Pavlick et al. 2015)

Discovering Stylistic Variations



Principal Component Analysis (PCA)

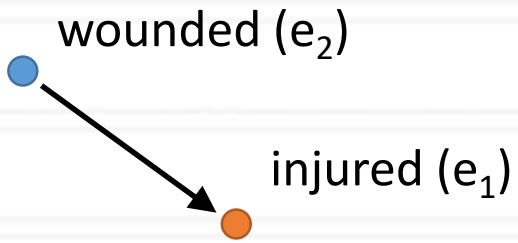
Discovering Stylistic Variations



1+ principal components (PCs)
→ stylistic subspace

Discovering Stylistic Variations – Method

- $D = \emptyset$
- For each lexical pair in PPDB:
  - Get embeddings (e_1, e_2)
  - $d = e_1 - e_2$
  - $D = D \cup \{d\}$
- PCA(D) → principle components



Discovering Stylistic Variations in Distributional Vector Space Models via Lexical Paraphrases

- How can we robustly discover stylistic variations in vector space?

- What does a stylistic subspace look like?

Stylistic Subspace: Top-k Principle Components

k	Representative paraphrases (largest projections)			
1	annulling – canceling programme – program organised – organize	abolished – canceled imperatives – essentials six-party – six-way	centre – center motorway – freeway tranquility – serenity	emphasise – highlight labour – labor tripartite – three-way
2	spendings – expenditures doctor – physician	summons – subpoenas falls – decreases	anti-malaria –antimalarial banned – prohibiting	
3	decreased – receded	decreased – fallen	decreased – declined	decreased – shrank
4	agreements – understandings discriminatory – discriminative	unlimited – unbounded timetable – time-scale	disruptions – perturbations amended – altered	
5	underscored – underline widened – broaden	eliminated – delete emphasises – underline	highlights – underline decreased – reduce	widened – expand performed – fulfil

word2vec on ICWSM 2009 Spinn3r
1.6 billion words from blogs

Stylistic Subspace: Top-k Principle Components

American/British-English

k	Representative paraphrases			
1	annulling – canceling **programme – program** **organised – organize**	abolished – canceled imperatives – essentials six-party – six-way	**centre – center** **motorway – freeway** tranquility – serenity	emphasise – highlight **labour – labor** tripartite – three-way
2	spendings – expenditures **doctor – physician**	**summons – subpoenas** falls – decreases	anti-malaria –antimalarial banned – prohibiting	
3	**decreased – receded**	**decreased – fallen**	**decreased – declined**	**decreased – shrank**
4	agreements – understandings discriminatory – discriminative	unlimited – unbounded timetable – time-scale	disruptions – perturbations amended – altered	
5	underscored – underline widened – broaden	**eliminated – delete** emphasises – underline	highlights – underline decreased – reduce	widened – expand performed – fulfil

subtle differences

formality variations

biased PPDB vocabulary

Stylistic Subspace: Top-k Principle Components

k	Representative paraphrases			
1	annulling – canceling programme – program organised – organize	abolished – canceled imperatives – essentials six-party – six-way	centre – center motorway – freeway tranquility – serenity	emphasise – highlight labour – labor tripartite – three-way
2	spendings – expenditures doctor – physician	summons – subpoenas falls – decreases	anti-malaria –antimalarial banned – prohibiting	
3	decreased – receded	decreased – fallen	decreased – declined	decreased – shrank
4	agreements – understandings discriminatory – discriminative	unlimited – unbounded timetable – time-scale	disruptions – perturbations amended – altered	
5	underscored – underline widened – broaden	eliminated – delete emphasises – underline	highlights – underline decreased – reduce	widened – expand performed – fulfil

- Signals of stylistic variations
- Difficult to align each PC to a clear-cut style dimension

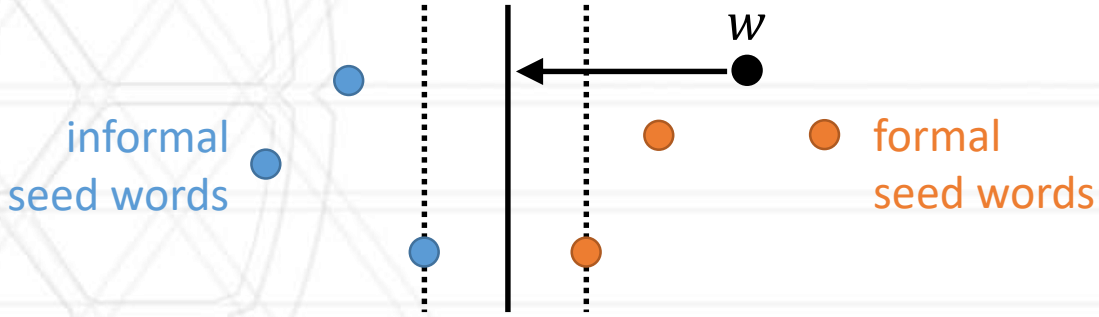
Discovering Stylistic Variations in Distributional Vector Space Models via Lexical Paraphrases

- How can we robustly discover stylistic variations in vector space?
- What does a stylistic subspace look like?
- Is subspace useful to detect stylistic variations?
  - How to use it? How many top PCs?

Quantitative Evaluation: Lexical Formality Scoring

- Task (Brooke et al. 2010)
  - Input: a synonym pair
    - enormous – huge
  - Output: a binary prediction
    - which one is more formal (enormous > huge)
- Training data
  - 138 informal seed words
  - 105 formal seed words
- Test data
  - 399 synonym pairs

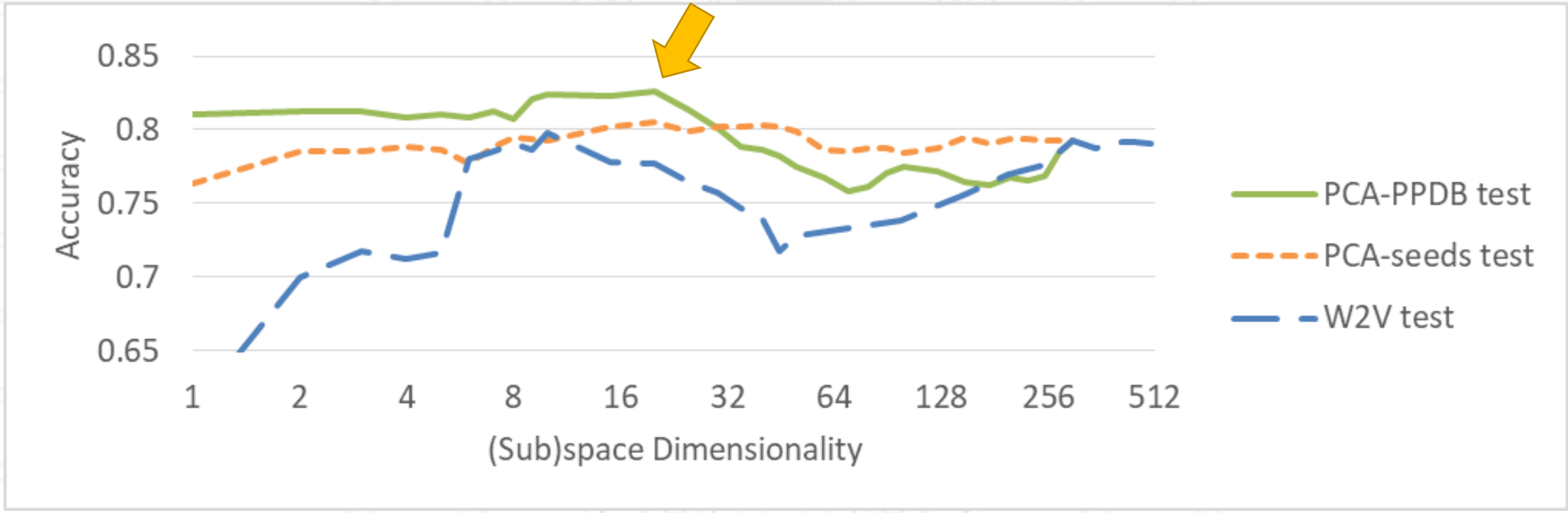
- Method
  - Linear SVM



Quantitative Evaluation: Lexical Formality Scoring

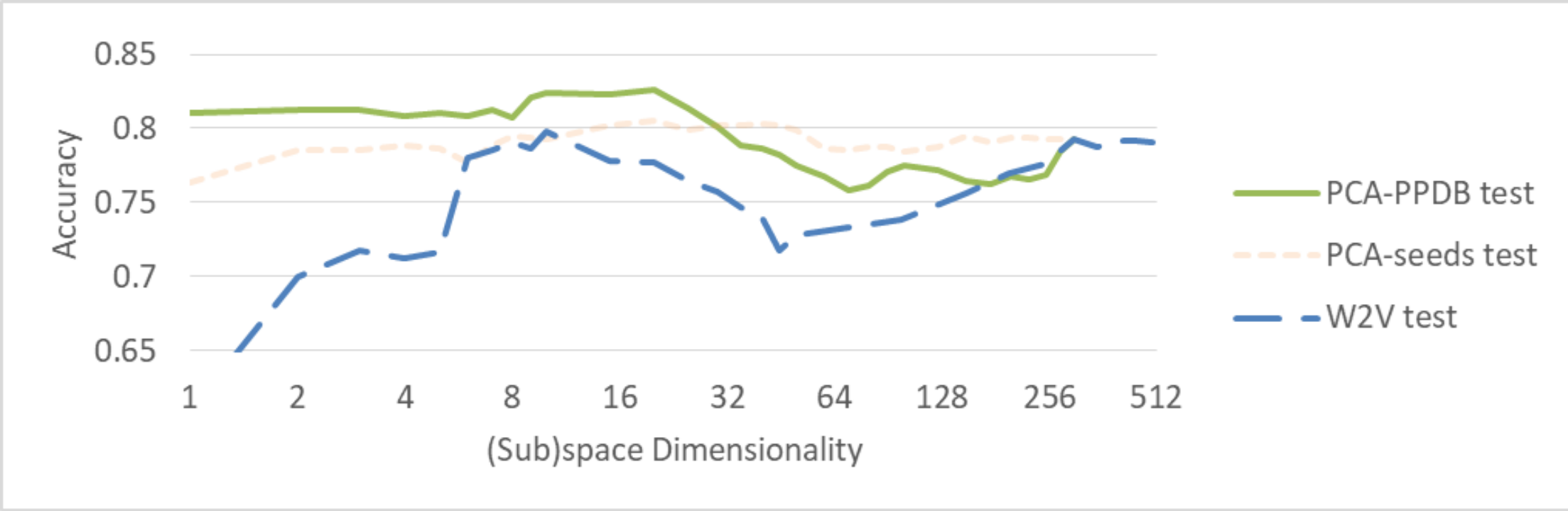
- Which subspace properties might impact formality scoring?
  - Dimensionality?
    - Word2vec (W2V): 1-500
    - PCA-PPDB: 1-300
  - Training data of PCA?
    - PCA-seeds: 1-300

Results – Impact of Dimensionality



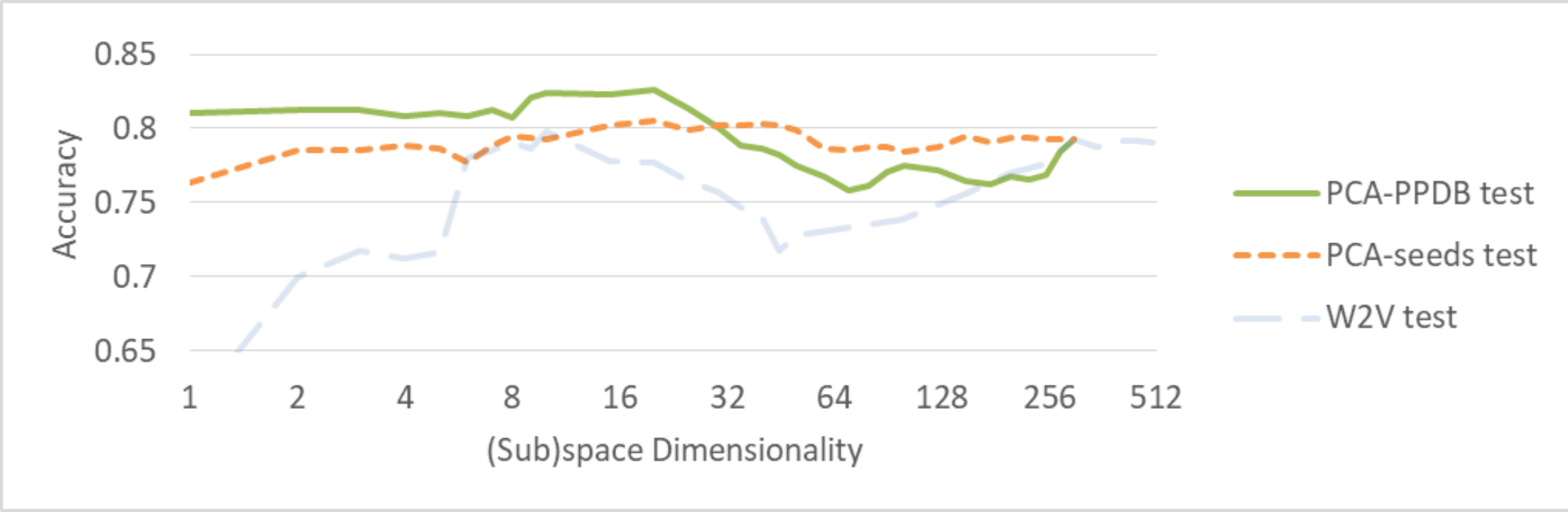
- Best: PCA-PPDB, Dimensionality=20 → Accuracy=0.826

Results – Stylistic Subspaces Yield Accurate Formality Predictions



- PCA-PPDB subspaces outperforms original word2vec spaces (baseline).

Analysis – Impact of Training Data



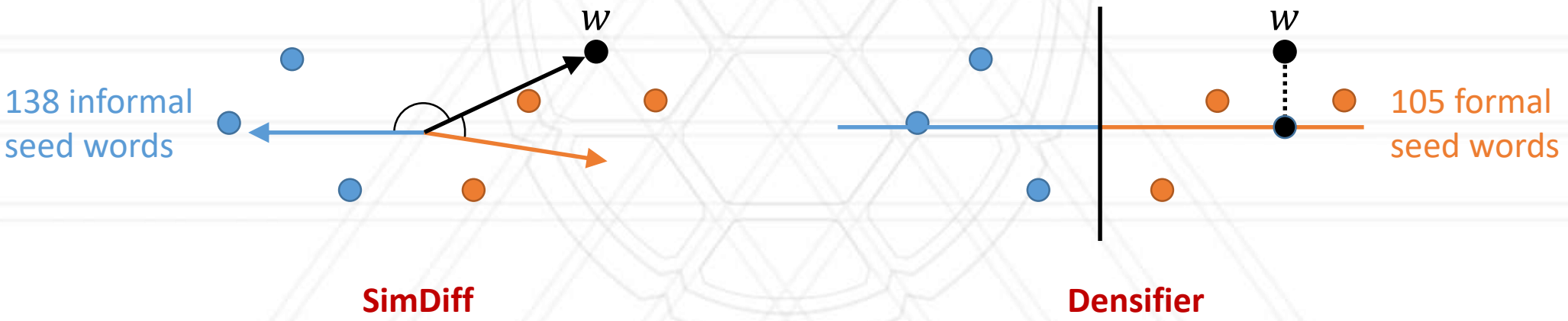
- PCA-seeds does not generalize as well as PCA-PPDB.

Other Formality Models

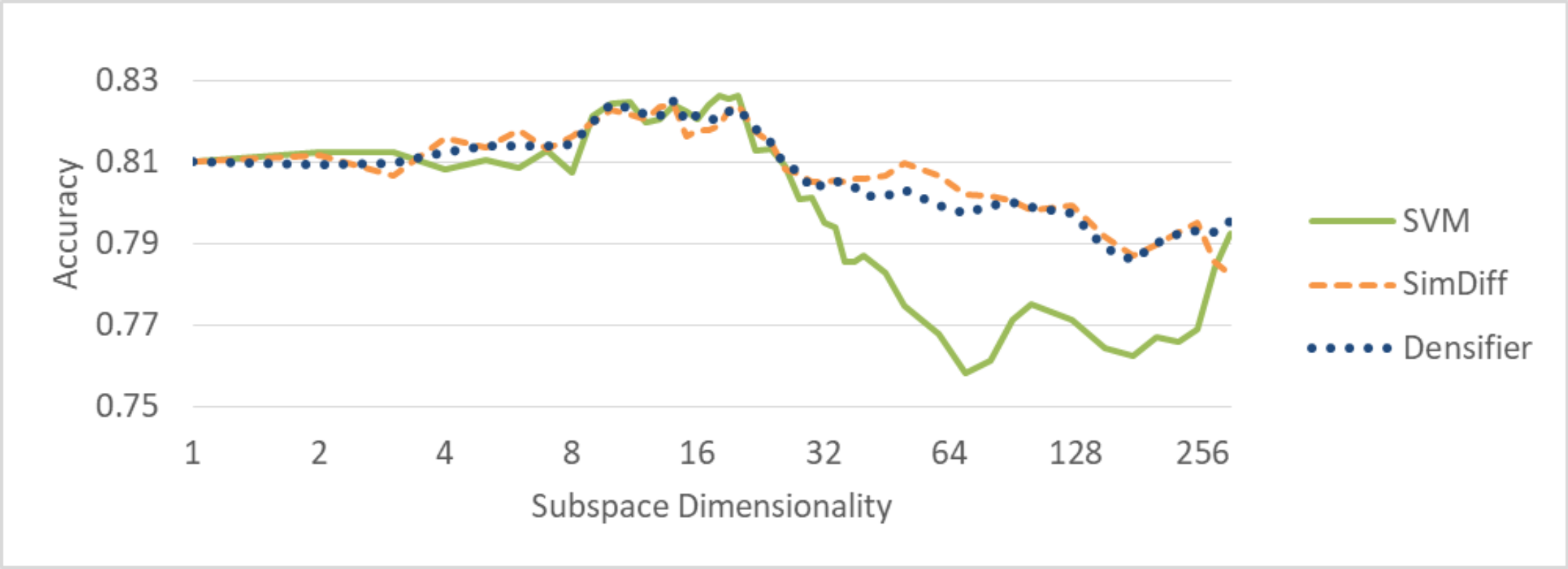
- Stylistic subspaces are effective for predicting formality using SVM.
- Factors:
  - Dimensionality? ✓
  - Training data of PCA? ✓
  - Can other more advanced methods perform even better than linear SVM?

Other Formality Models

- Can other more advanced methods perform even better?
  - **SimDiff** (Brooke et al. 2010) compares words to formal vs. informal seeds.
  - **Densifier** (Rothe et al. 2016) optimizes a formality dimension that aims at separating formal/informal words and grouping words in the same set.



Results – Different Methods (with PCA-PPDB)



- All methods using the same subspace perform similarly.

Discovering Stylistic Variations in Distributional Vector Space Models via Lexical Paraphrases

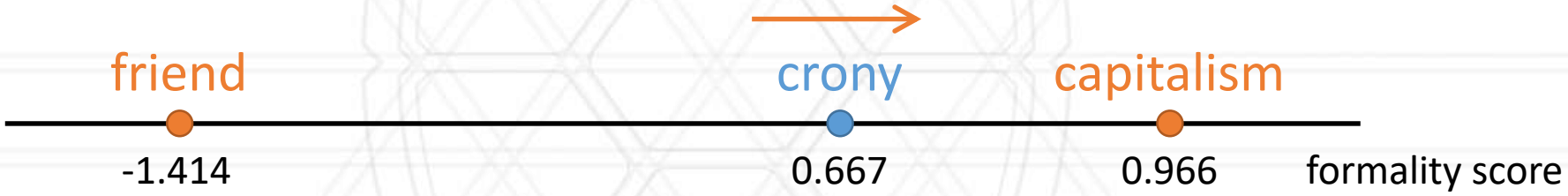
- How can we robustly discover stylistic variations in vector space?
- What does a stylistic subspace look like?
- Is subspace useful to detect stylistic variations?
- What kinds of prediction errors remain?

Error Analysis (with PCA-PPDB, Dim=20, SVM)

- What kinds of prediction errors remain?

Error Analysis

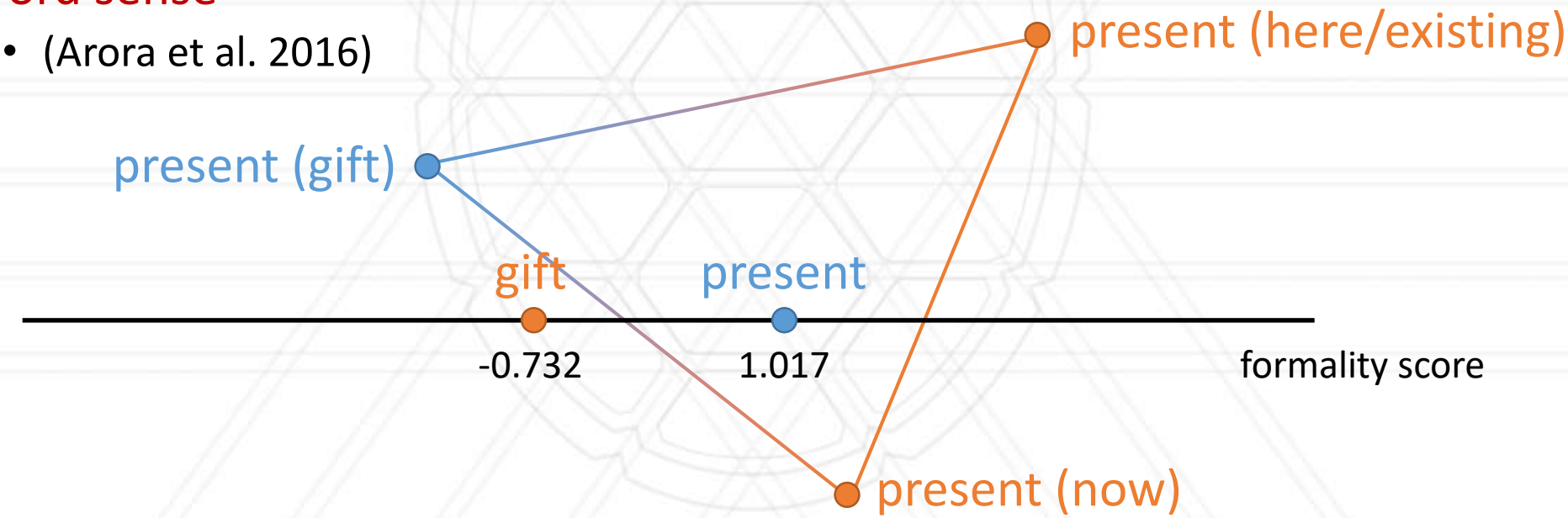
- What kinds of prediction errors remain?
  - Word association



  - "crony capitalism": mutually advantageous relationships between business leaders and government officials

Error Analysis

- What kinds of prediction errors remain?
  - Word association
  - Word sense
    - (Arora et al. 2016)



Error Analysis

- What kinds of prediction errors remain?
  - Word association
  - Word sense
  - Word frequency
    - more-frequent-less-formal
    - good predictor in this dataset
      - 0.771

correct predictions

less formal words	more formal words
grill (10x)	interrogate
excuse (10x)	remit
gardening (100x)	tillage
get (100x)	obtain
hurry (10x)	expedite
catch (100x)	apprehend
watch (10x)	observe
loud (100x)	clamorous
quote (1000x)	adduce
beach (100x)	littoral

Error Analysis

- What kinds of prediction errors remain?
  - Word association
  - Word sense
  - Word frequency
    - more-frequent-less-formal ✗
    - biased the prediction
    - frequency can be partially reconstructed (Rothe et al. 2016)

incorrect predictions

less formal words	more formal words
crony	friend (100x)
conceit	vanity (10x)
present (1x)	gift
shiv	knife (10x)
quotation	quote (10x)
frighten	scare (10x)
phony	fake (1x)
parched	dehydrated (1x)
punish (10x)	chasen
penetrating (10x)	perspicacious

Discovering Stylistic Variations in Distributional Vector Space Models via Lexical Paraphrases

- How can we robustly discover stylistic variations in vector space?
  - our proposal: PCA + paraphrases → a stylist subspace
- What does a stylistic subspace look like?
  - qualitative analysis: some evidences of stylistic variations in PC, but no clear-cut mapping from PC to style
- Is subspace useful to detect stylistic variations?
  - quantitative evaluation on lexical formality task
    - Inducing a stylistic subspace helps detect formality
    - Which factors Impact scoring? Dimensionality ✓, training data ✓, prediction methods ✗
- What kinds of prediction errors remain?
  - error analysis: limitations of using vector space models

Code: https://github.com/xingniu/computational-stylistic-variations

DEPARTMENT OF
COMPUTER SCIENCE