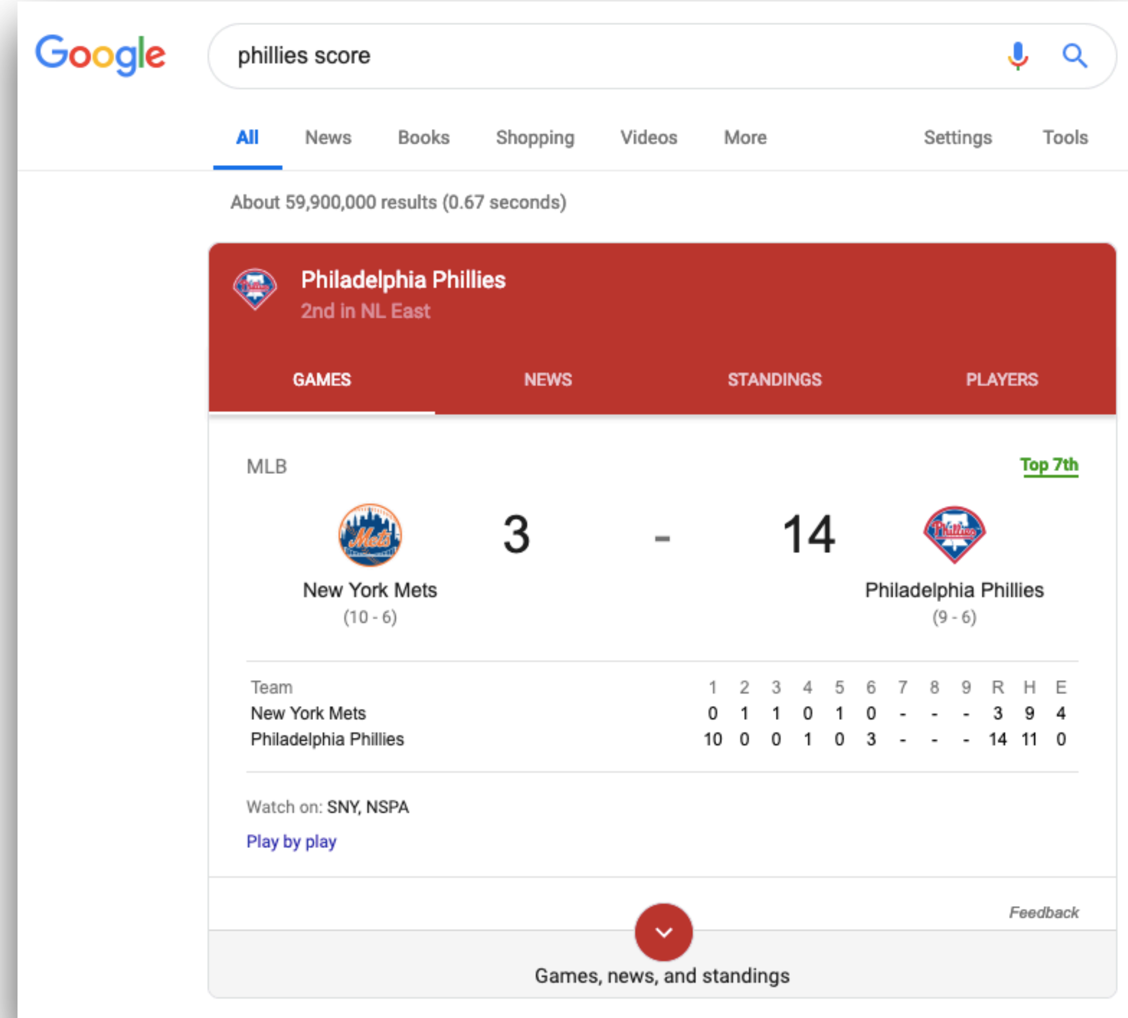
Paraphrase-Based Models of Lexical Semantics

Dissertation Defense

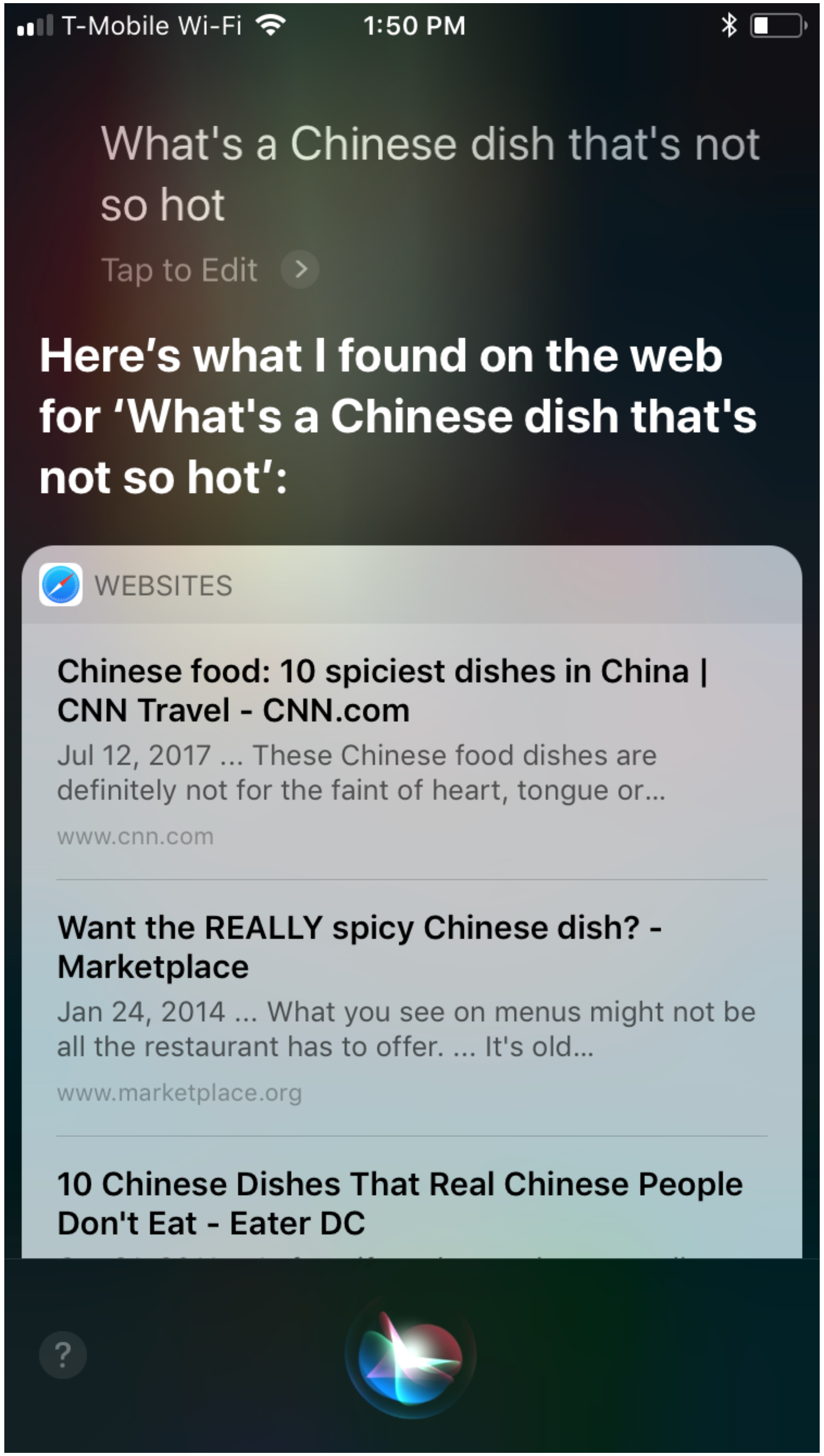
Anne Cocos
Department of Computer and Information Science
University of Pennsylvania


Google
phillies score
All
News
Books
Shopping
Videos
More
Settings
Tools
About 59,900,000 results (0.67 seconds)
Philadelphia Phillies
2nd in NL East
GAMES
NEWS
STANDINGS
PLAYERS
MLB
Top 7th
3 - 14
New York Mets
(10 - 6)
Philadelphia Phillies
(9 - 6)
Team 1 2 3 4 5 6 7 8 9 R H E
New York Mets 0 1 1 0 1 0 - - - 3 9 4
Philadelphia Phillies 10 0 0 1 0 3 - - - 14 11 0
Watch on: SNY, NSPA
Play by play
Feedback
Games, news, and standings


amazon

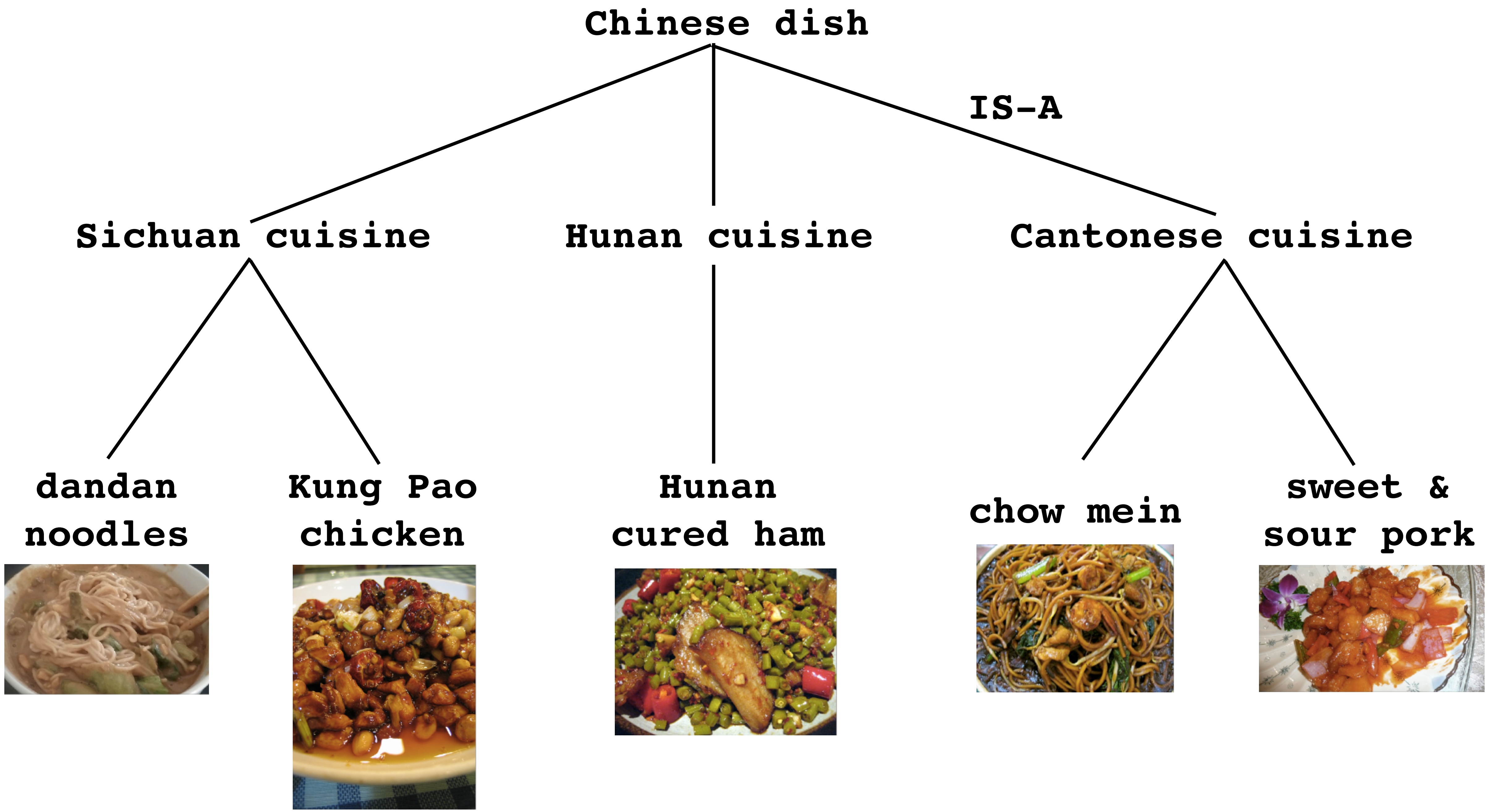

11:20
Hey Siri where are you
Tap to Edit
Wherever you are, that's where I am.

"What's a Chinese dish that's not so hot?"


T-Mobile Wi-Fi
1:50 PM
What's a Chinese dish that's not so hot
Tap to Edit
Here's what I found on the web for 'What's a Chinese dish that's not so hot':
WEBSITES
Chinese food: 10 spiciest dishes in China | CNN Travel - CNN.com
Jul 12, 2017 ... These Chinese food dishes are definitely not for the faint of heart, tongue or...
www.cnn.com
Want the REALLY spicy Chinese dish? - Marketplace
Jan 24, 2014 ... What you see on menus might not be all the restaurant has to offer. ... It's old...
www.marketplace.org
10 Chinese Dishes That Real Chinese People Don't Eat - Eater DC
?

"What's a Chinese dish that's not so hot?"

"What's a Chinese dish that's not so hot?"



"What's a Chinese **dish** that's not so **hot**?"

"What's a Chinese **dish** that's not so **hot**?"

hot dish?

"What's a Chinese **dish** that's not so **hot**?"

hot dish?

"What's a Chinese **dish** that's not so **hot**?"

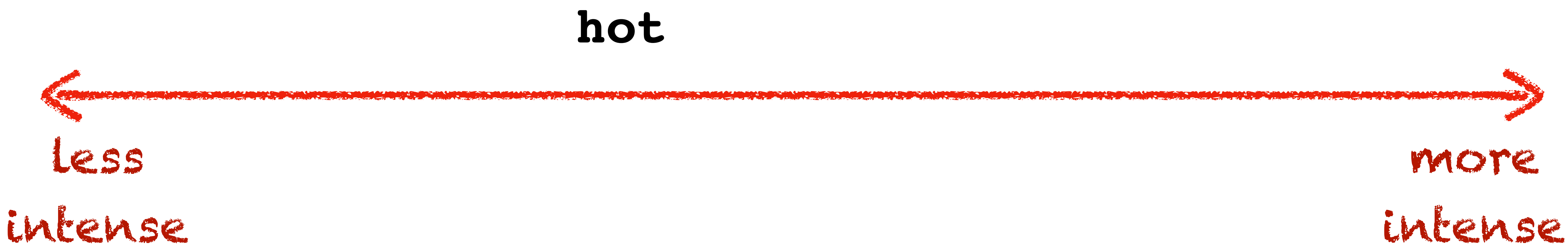
hot dish?

"What's a Chinese **dish** that's not so **hot**?"

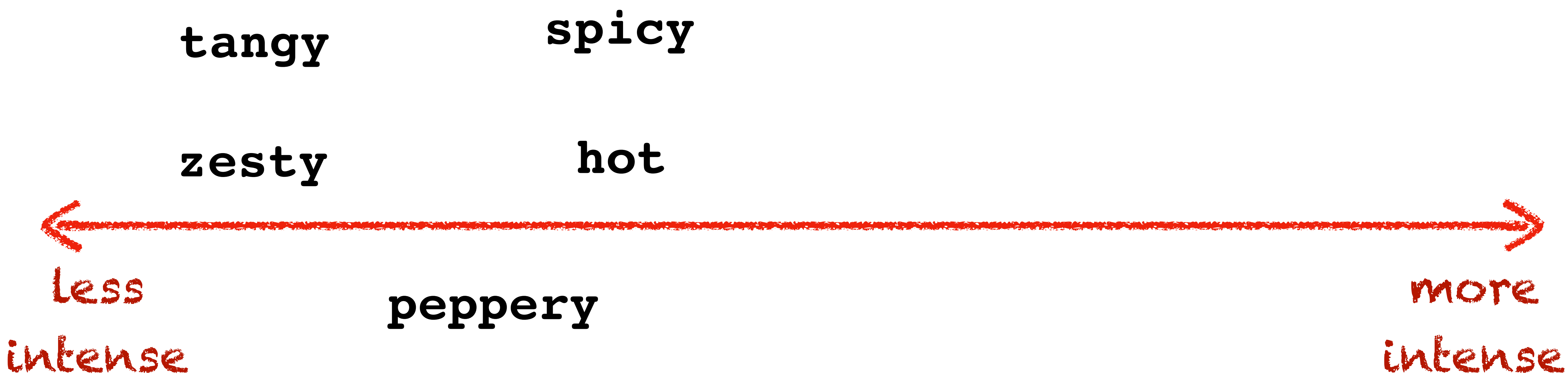
hot dish?

"What's a Chinese dish that's **not so hot**?"

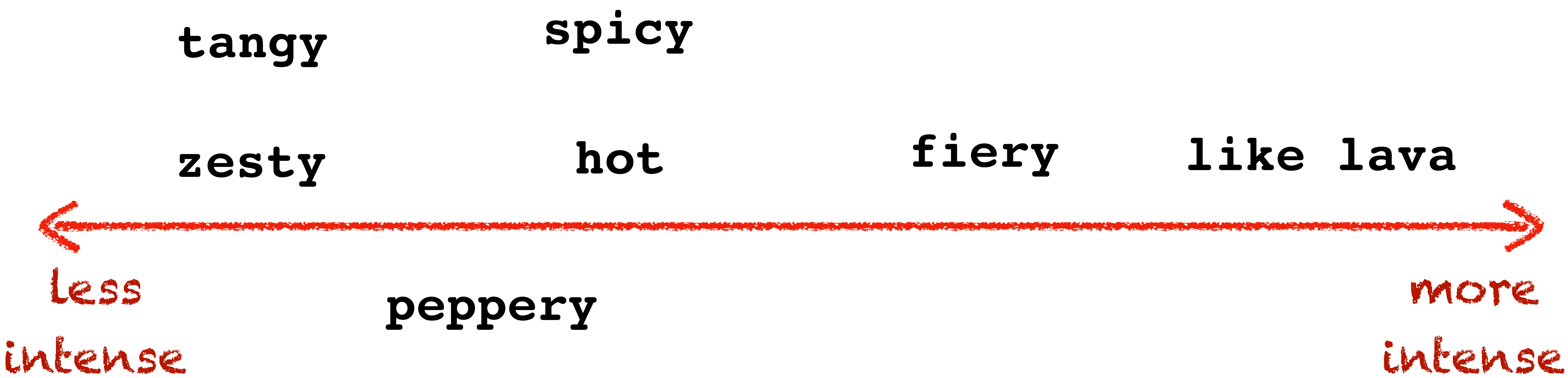
"What's a Chinese dish that's **not so hot**?"



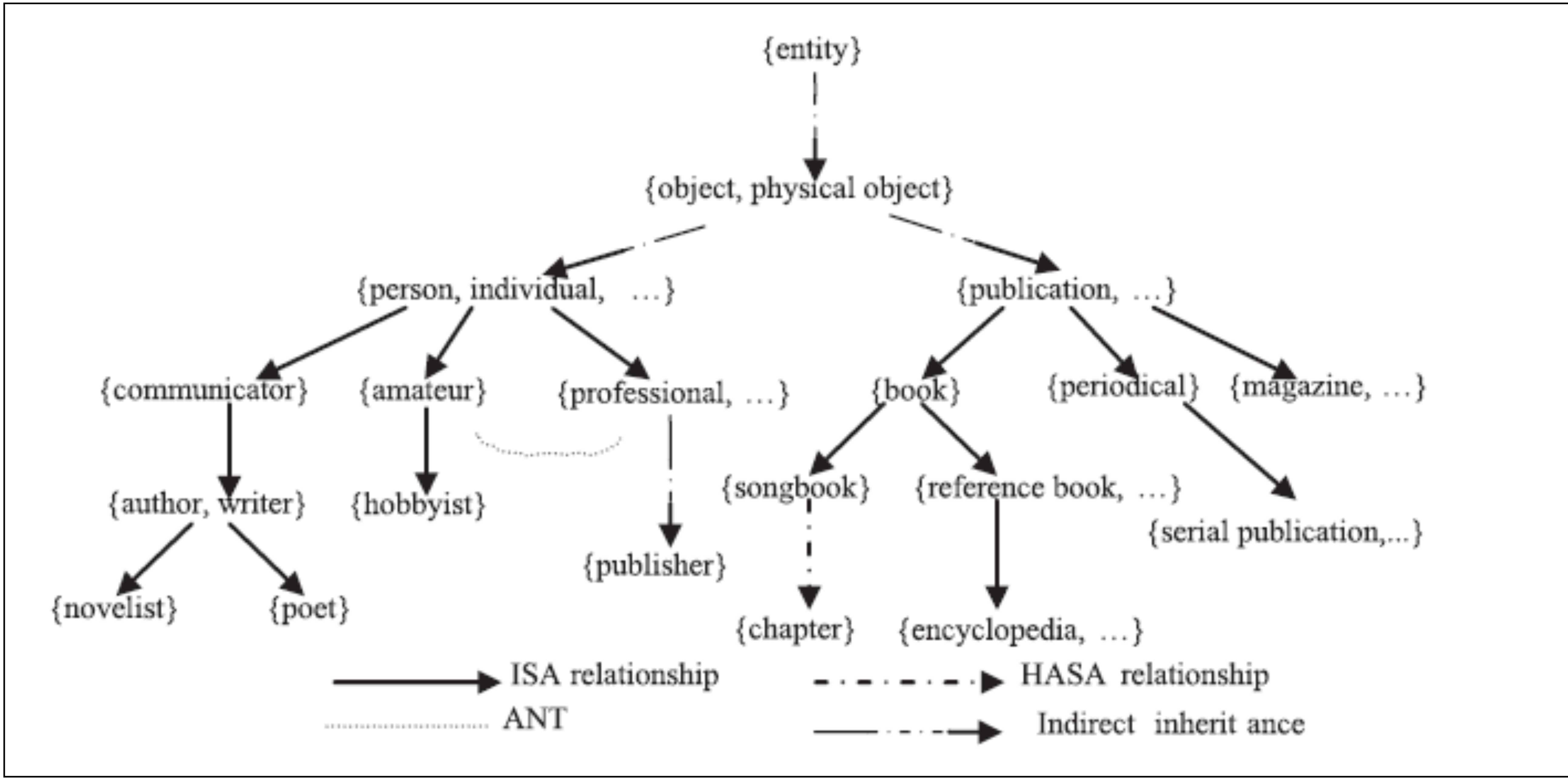
"What's a Chinese dish that's **not so hot**?"



“What’s a Chinese dish that’s **not so hot**?”

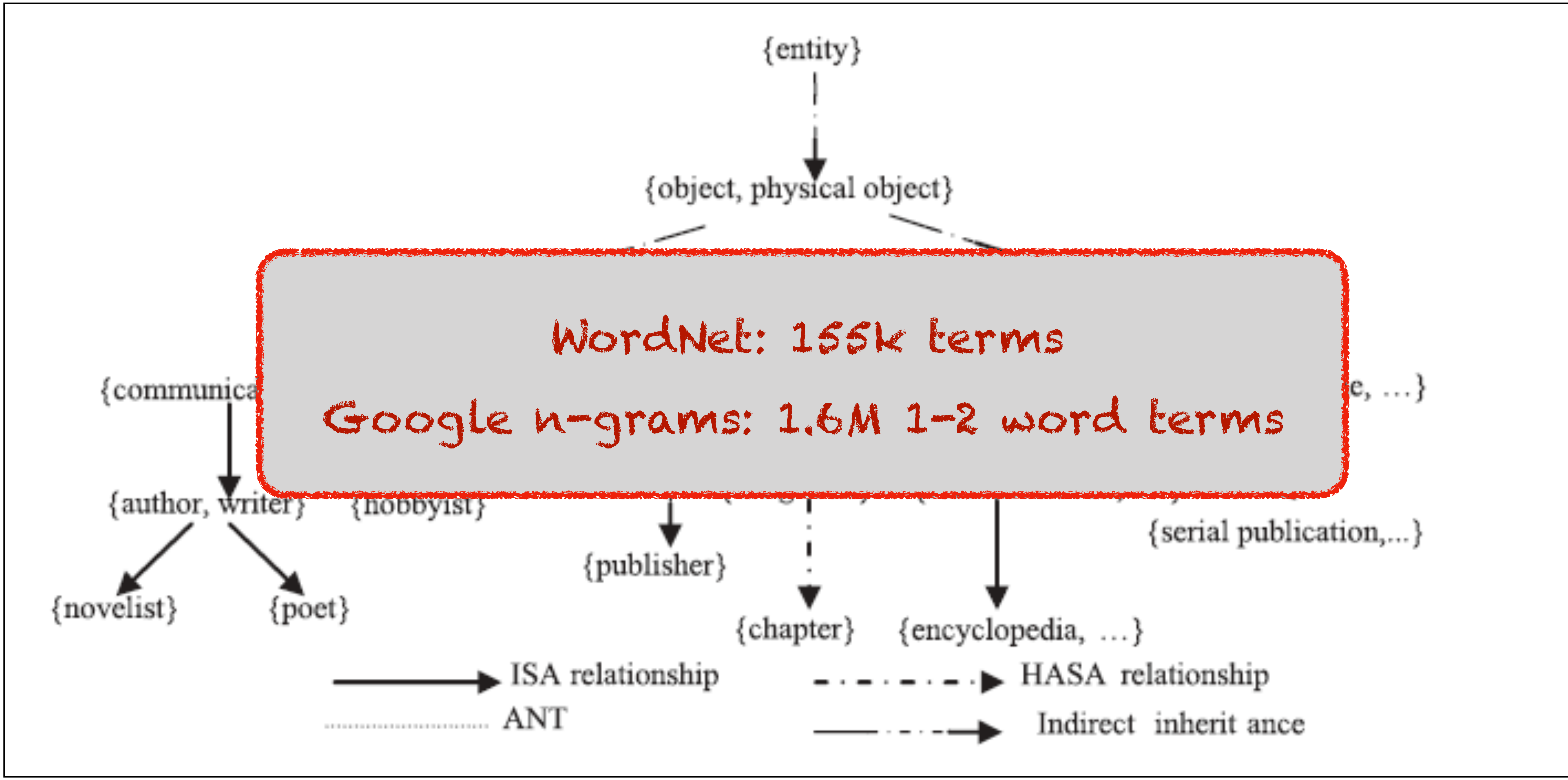


Semantic knowledge can be modeled **manually**.



WordNet (https://wordnet.princeton.edu/)

Semantic knowledge can be modeled **manually**.

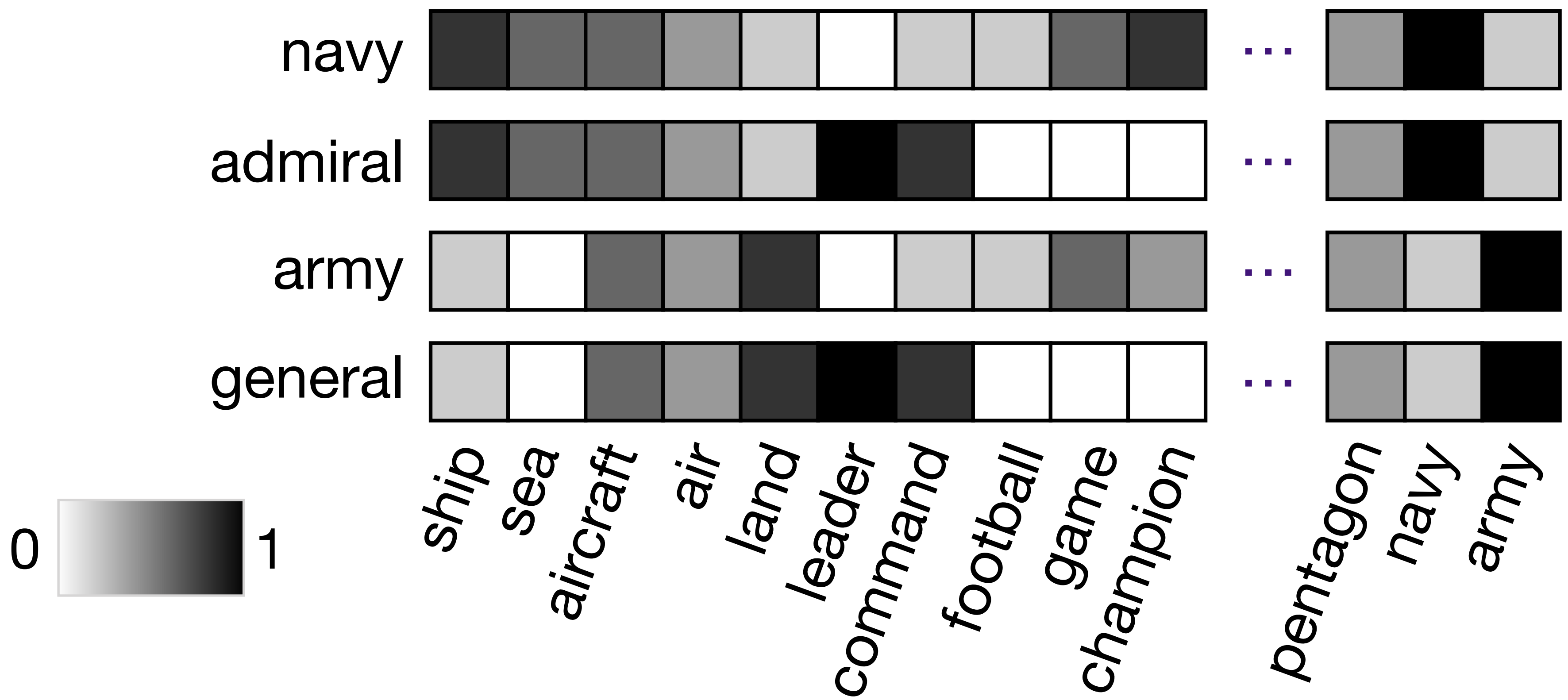


WordNet: 155k terms

Google n-grams: 1.6M 1-2 word terms

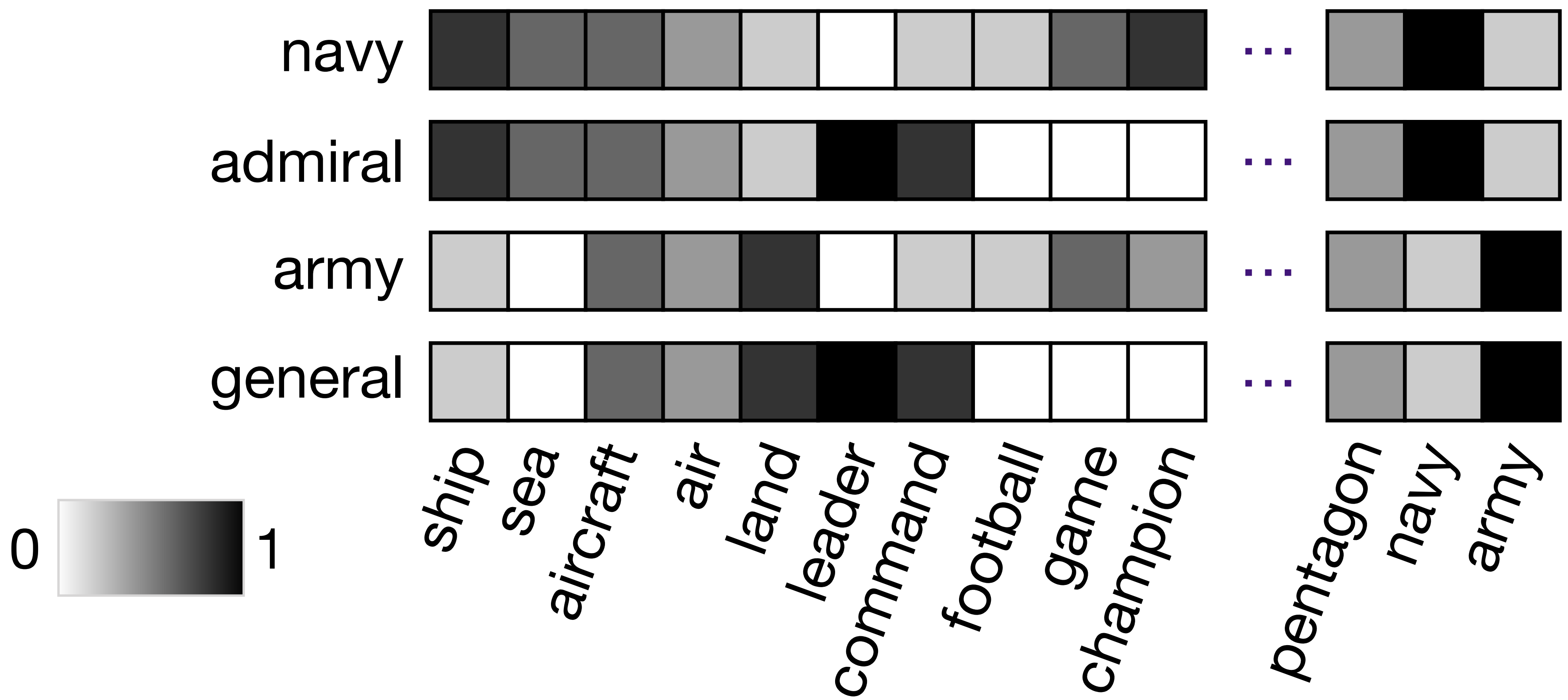
WordNet (https://wordnet.princeton.edu/)

Semantic knowledge can be modeled **automatically**.



Contextual Similarity Methods

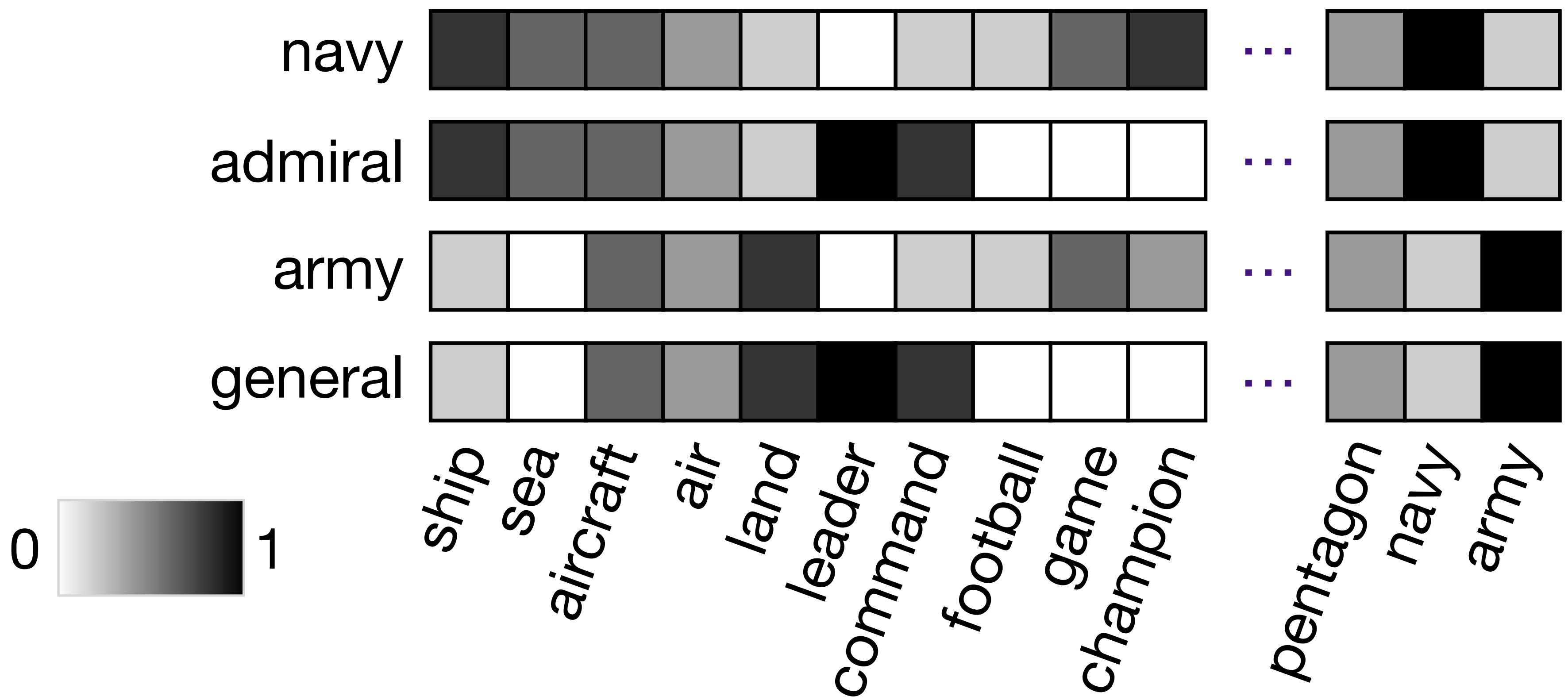
Semantic knowledge can be modeled **automatically**.



Contextual Similarity Methods

antonyms?

Semantic knowledge can be modeled **automatically**.

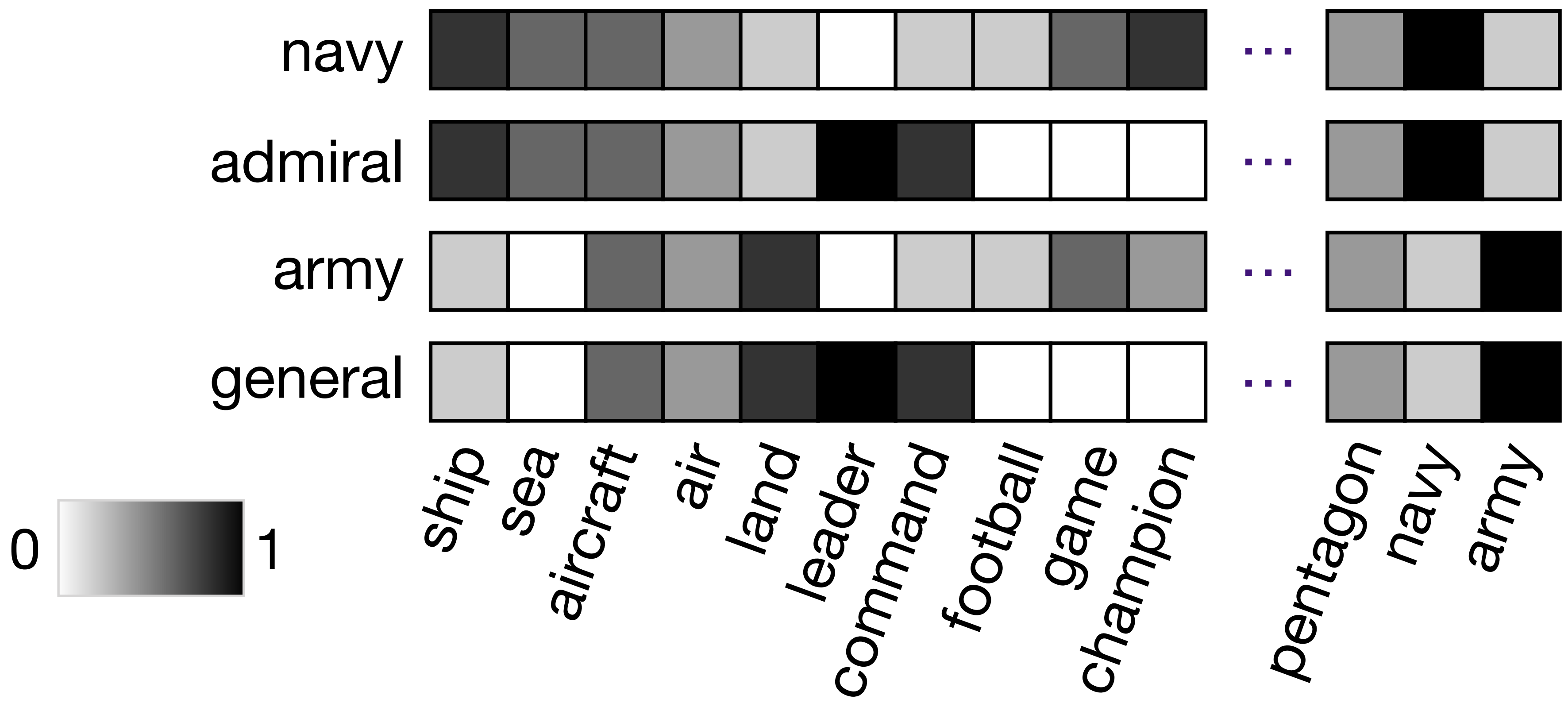


Contextual Similarity Methods

antonyms?

```
word2vec.similarity('hot','sizzling') = 0.51
word2vec.similarity('hot','cold')     = 0.48
word2vec.similarity('hot','steaming') = 0.45
```

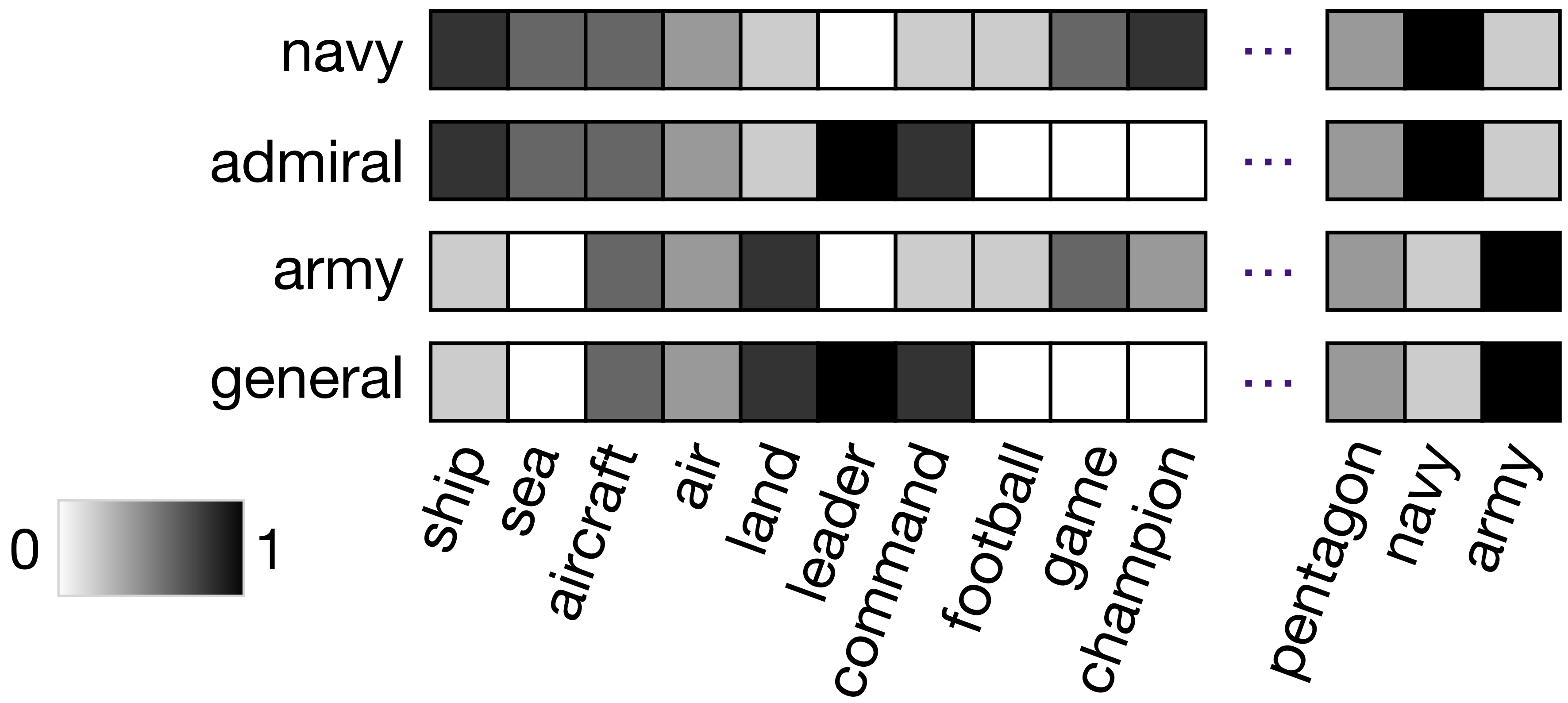
Semantic knowledge can be modeled **automatically**.



Contextual Similarity Methods

antonyms?

Semantic knowledge can be modeled **automatically**.

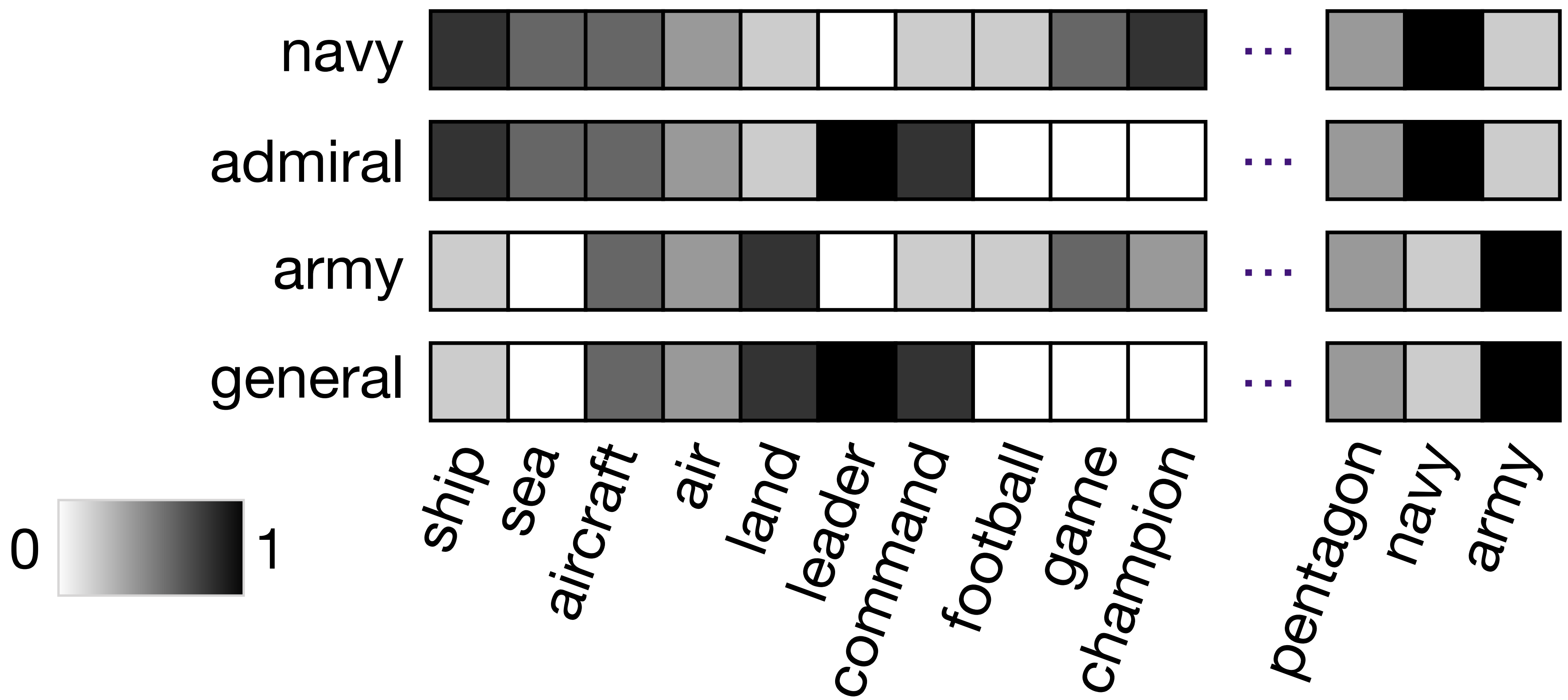


Contextual Similarity Methods

antonyms?

infrequent senses?

Semantic knowledge can be modeled **automatically**.



Contextual Similarity Methods

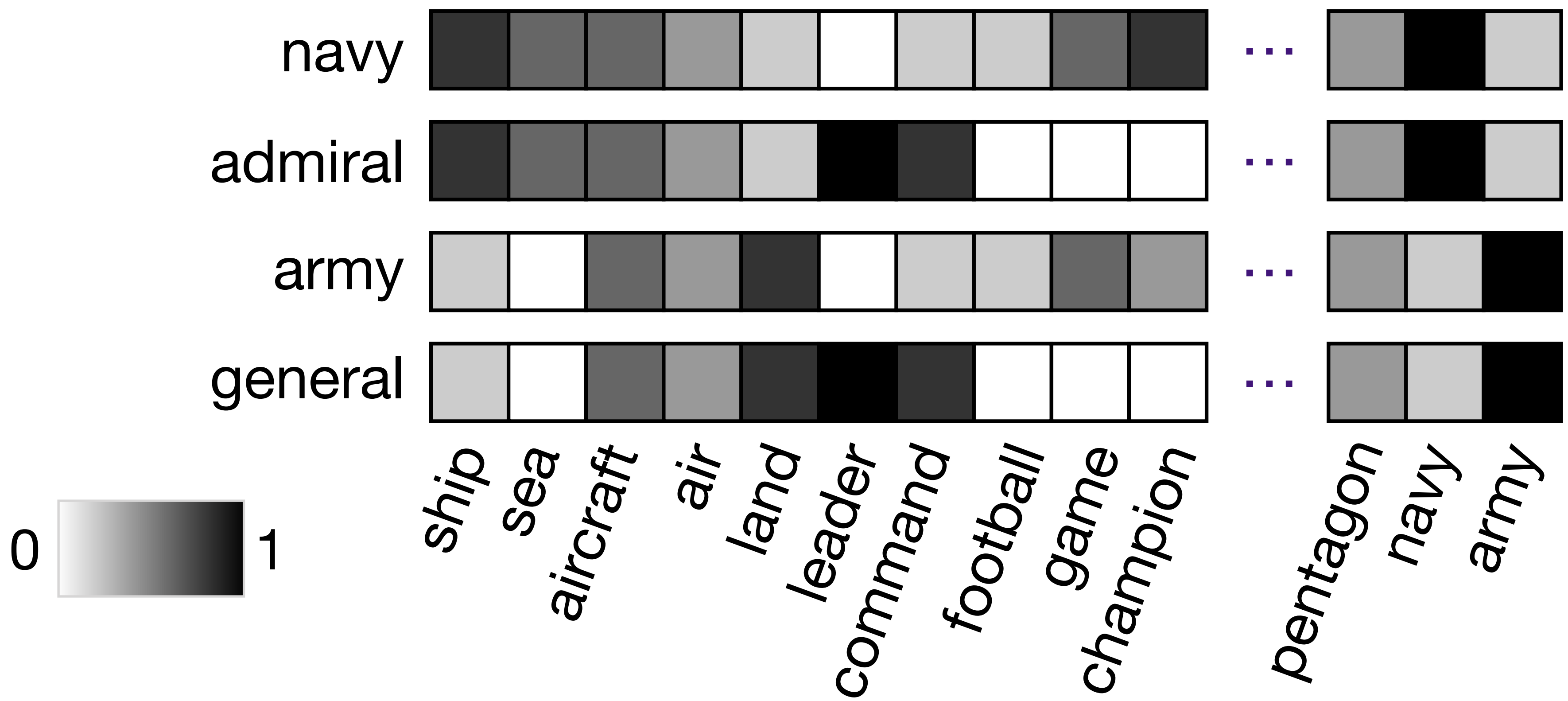
```
> word2vec.most_similar('crane')

cranes
cherry-picker
barge
scaffolding
9-ton
backhoe
excavator
forklift
14-ton
30-ton
```

antonyms?

infrequent senses?

Semantic knowledge can be modeled **automatically**.



Contextual Similarity Methods

```
> word2vec.most_similar('crane')

cranes
cherry-picker
barge
scaffolding
9-ton
backhoe
excavator
forklift
14-ton
30-ton
```

antonyms?

infrequent senses?

Semantic knowledge can be modeled **automatically**.

Lexico-Syntactic
Pattern Methods

[birds], such as [pigeons]

`pigeon IS-A bird`

not [great], but still [good]

`good < great`

Semantic knowledge can be modeled **automatically**.

Lexico-Syntactic Pattern Methods

[birds], such as [pigeons]

pigeon IS-A bird

not [great], but still [good]

good < great

synonyms?

Semantic knowledge can be modeled **automatically**.

Lexico-Syntactic Pattern Methods

[birds], such as [pigeons]

`pigeon IS-A bird`

not [great], but still [good]

`good < great`

synonyms?

which meaning? great [QUALITY] vs. great [SIZE]

My work aims to model semantic knowledge using **paraphrases.**

My work aims to model semantic knowledge using **paraphrases.**

Differing textual expressions of the same meaning:

My work aims to model semantic knowledge using **paraphrases.**

Differing textual expressions of the same meaning:

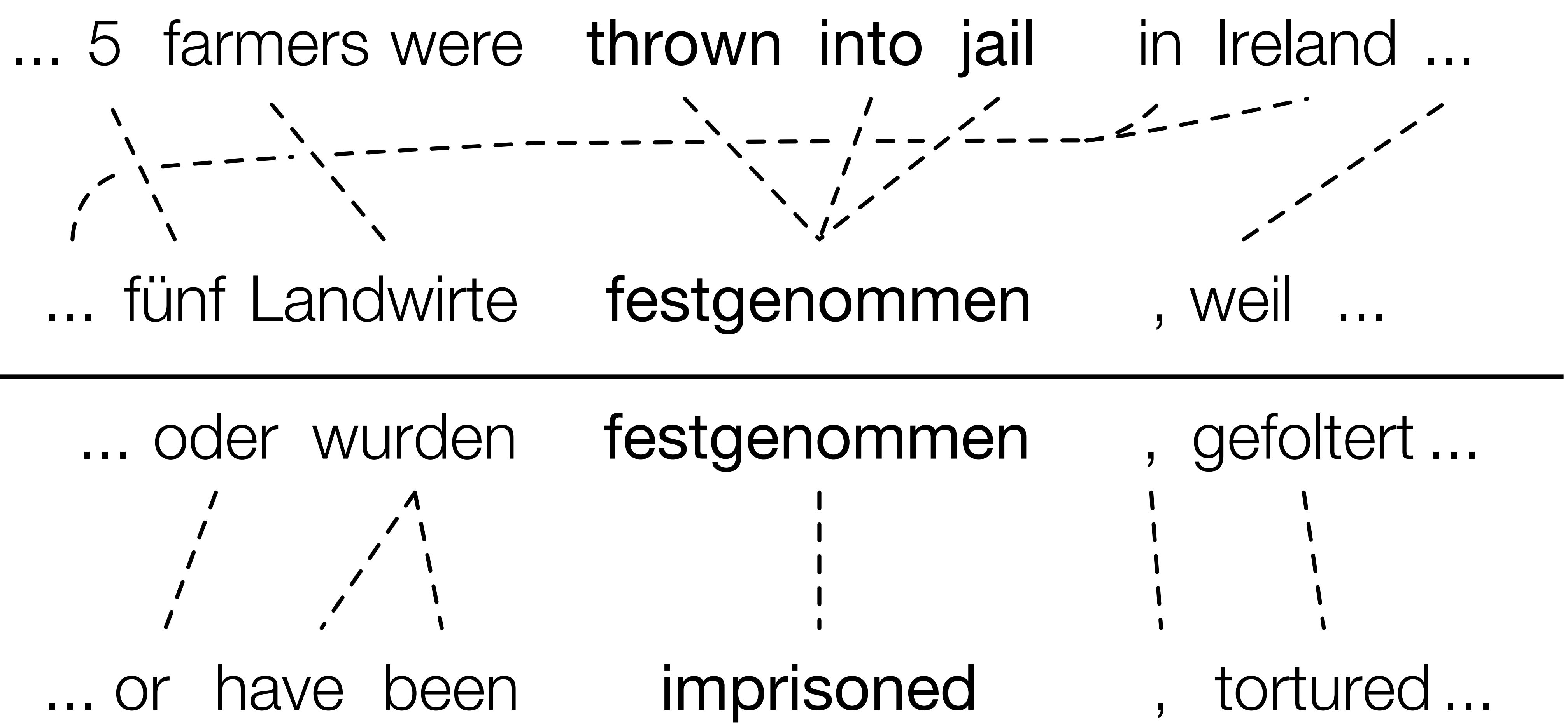
cup	$\leftrightarrow$	mug
the king's speech	$\leftrightarrow$	His Majesty's address
X_1 devours X_2	$\leftrightarrow$	X_2 is eaten by X_1
really tasty	$\leftrightarrow$	exquisite

My work aims to model semantic knowledge using paraphrases (acquired by bilingual pivoting)

Differing textual expressions of the same meaning:

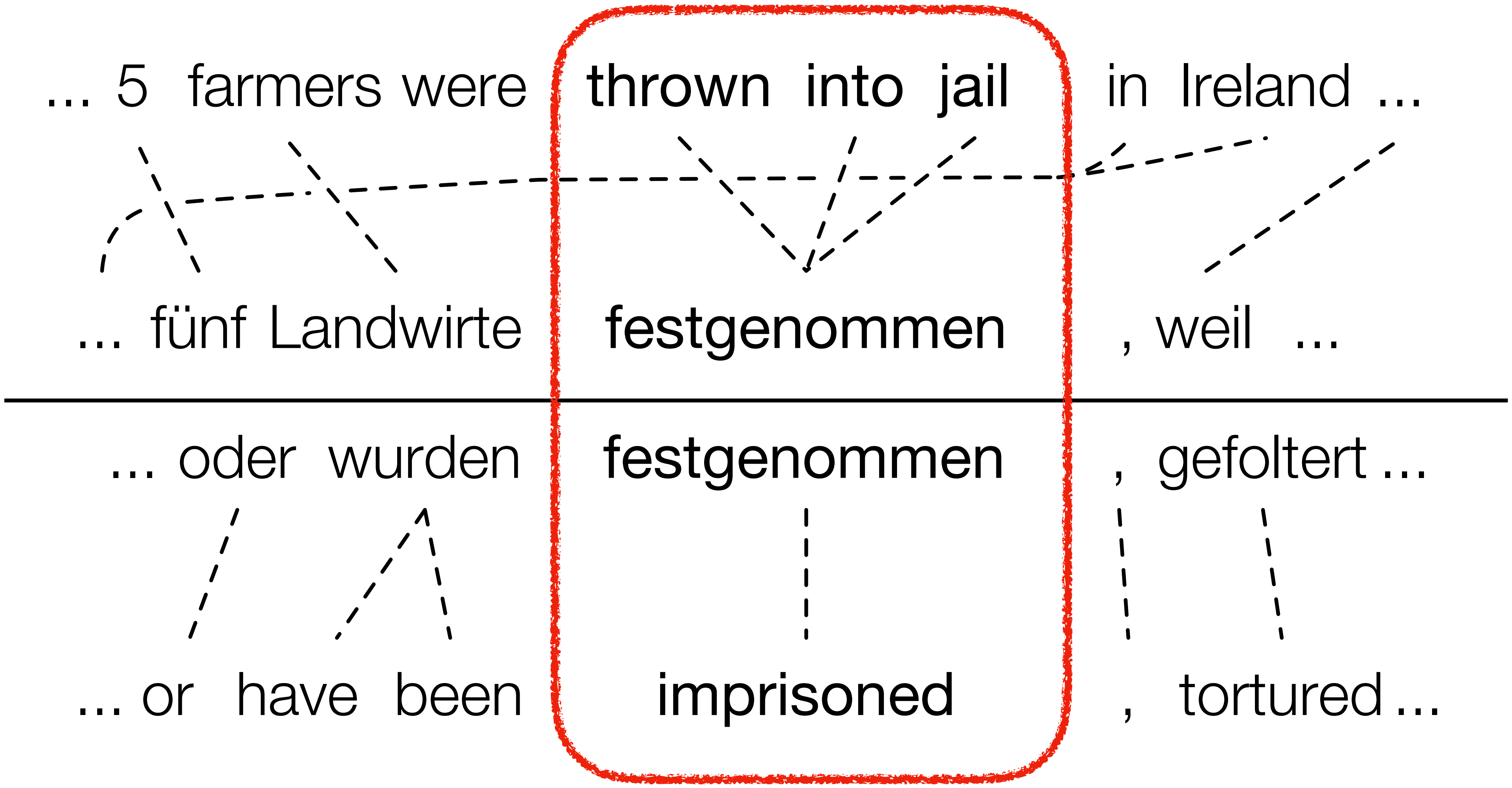
cup	↔	mug
the king's speech	↔	His Majesty's address
X_1 devours X_2	↔	X_2 is eaten by X_1
really tasty	↔	exquisite

Bilingual Pivoting



Paraphrasing with Bilingual Parallel Corpora. Colin Bannard and Chris Callison-Burch. ACL 2005.

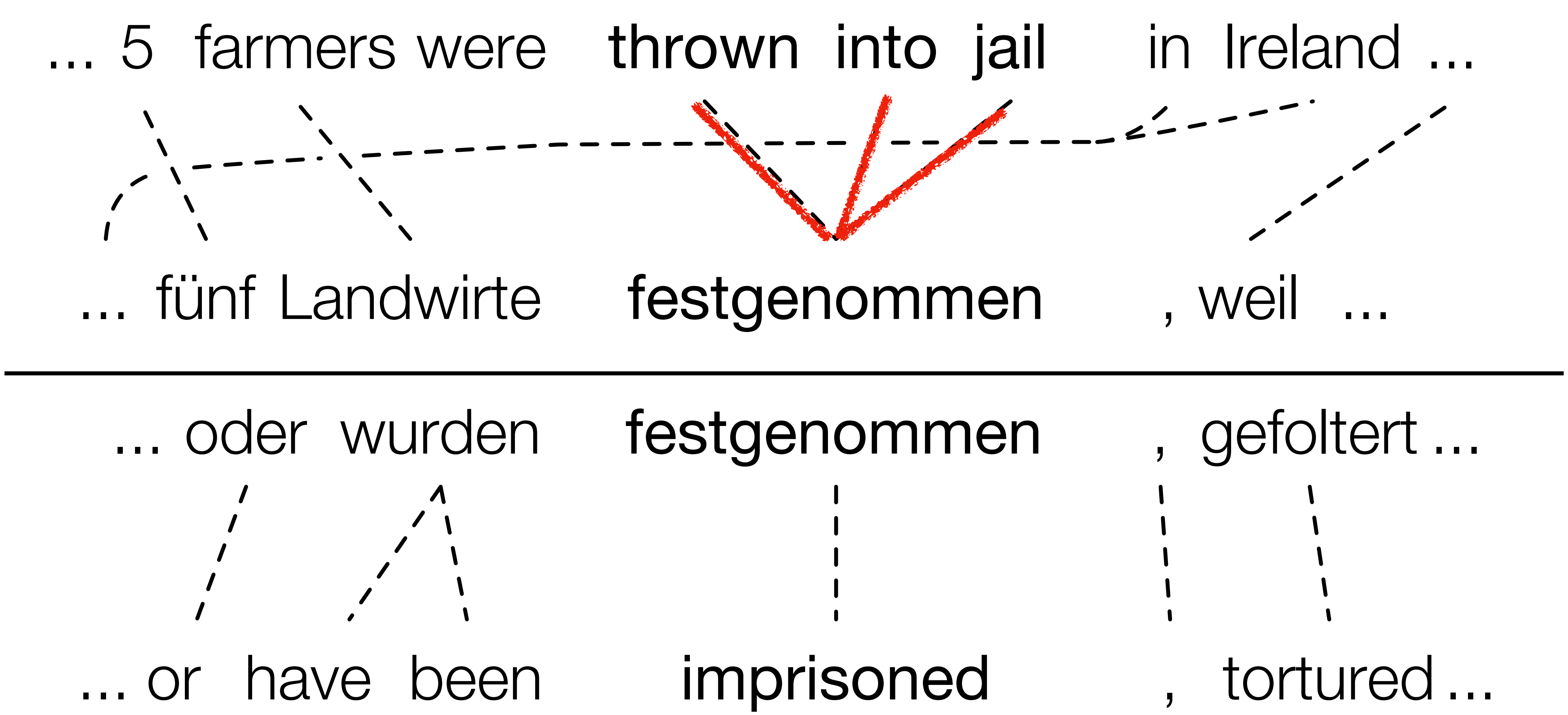
Bilingual Pivoting



Paraphrasing with Bilingual Parallel Corpora. Colin Bannard and Chris Callison-Burch. ACL 2005.

Bilingual Pivoting

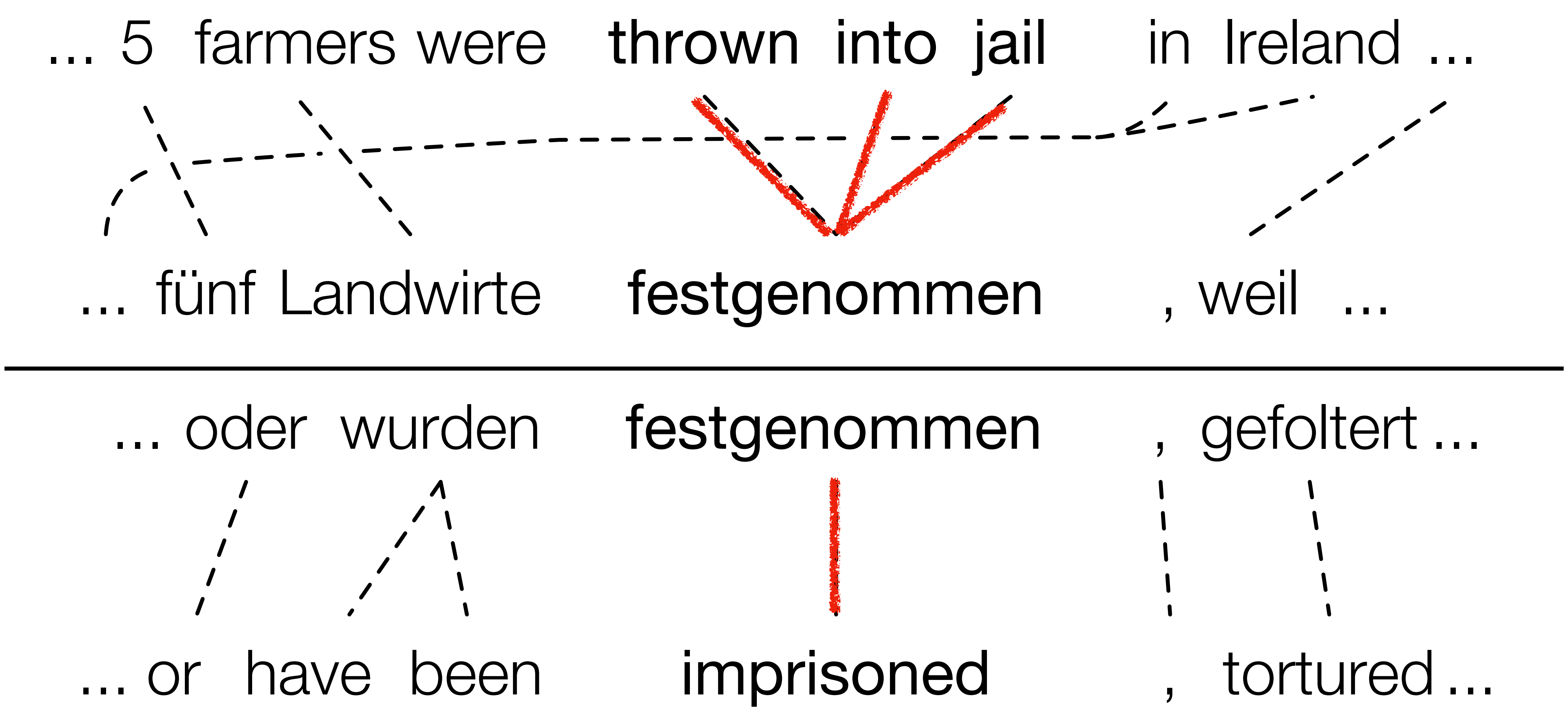
p(*"thrown into jail"* | *"festgenommen"*)



Paraphrasing with Bilingual Parallel Corpora. Colin Bannard and Chris Callison-Burch. ACL 2005.

Bilingual Pivoting

p("*thrown into jail*" | "*festgenommen*")



p("*festgenommen*" | "*imprisoned*")

Paraphrasing with Bilingual Parallel Corpora. Colin Bannard and Chris Callison-Burch. ACL 2005.

Bilingual Pivoting

p(*"thrown into jail"* | *"festgenommen"*)

p(*"festgenommen"* | *"imprisoned"*)

Paraphrasing with Bilingual Parallel Corpora. Colin Bannard and Chris Callison-Burch. ACL 2005.

Bilingual Pivoting

p(*"thrown into jail"* | *"festgenommen"*)

p(*"festgenommen"* | *"imprisoned"*)

Paraphrasing with Bilingual Parallel Corpora. Colin Bannard and Chris Callison-Burch. ACL 2005.

Bilingual Pivoting

$\boxed{\text{p(“}\textit{thrown into jail}\text{” | “}\textit{festgenommen}\text{”)}} = p(e_1 \mid f)$

$\boxed{\text{p(“}\textit{festgenommen}\text{” | “}\textit{imprisoned}\text{”)}} = p(f \mid e_2)$

Paraphrasing with Bilingual Parallel Corpora. Colin Bannard and Chris Callison-Burch. ACL 2005.

Bilingual Pivoting

$\boxed{\text{p(“\textit{thrown into jail}” | “\textit{festgenommen}”)}} = p(e_1 \mid f)$

$\boxed{\text{p(“\textit{festgenommen}” | “\textit{imprisoned}”)}} = p(f \mid e_2)$

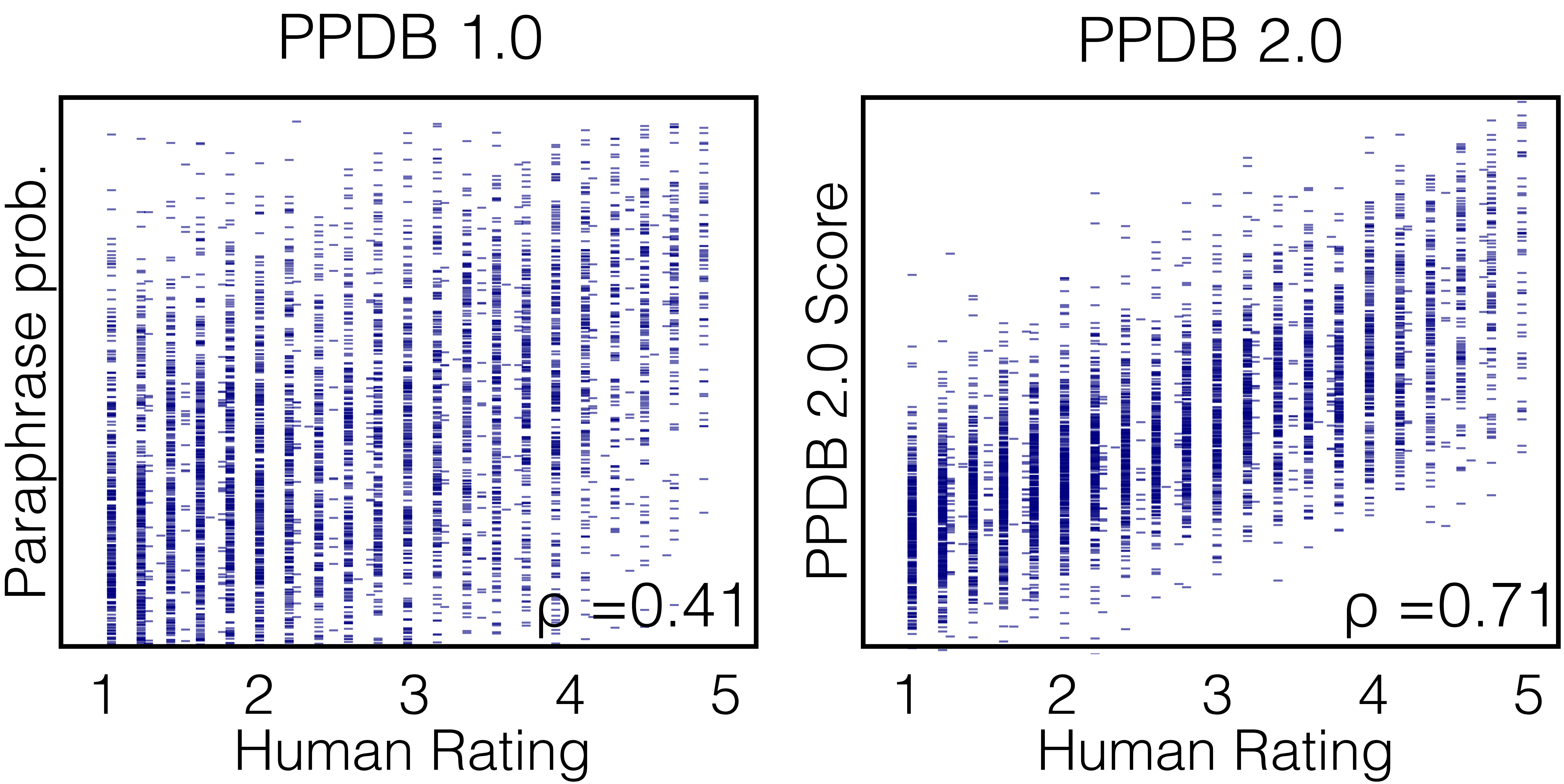
$$p(e_1 \mid e_2) \approx \sum_f p(e_1 \mid f) \cdot p(f \mid e_2)$$

paraphrase probability

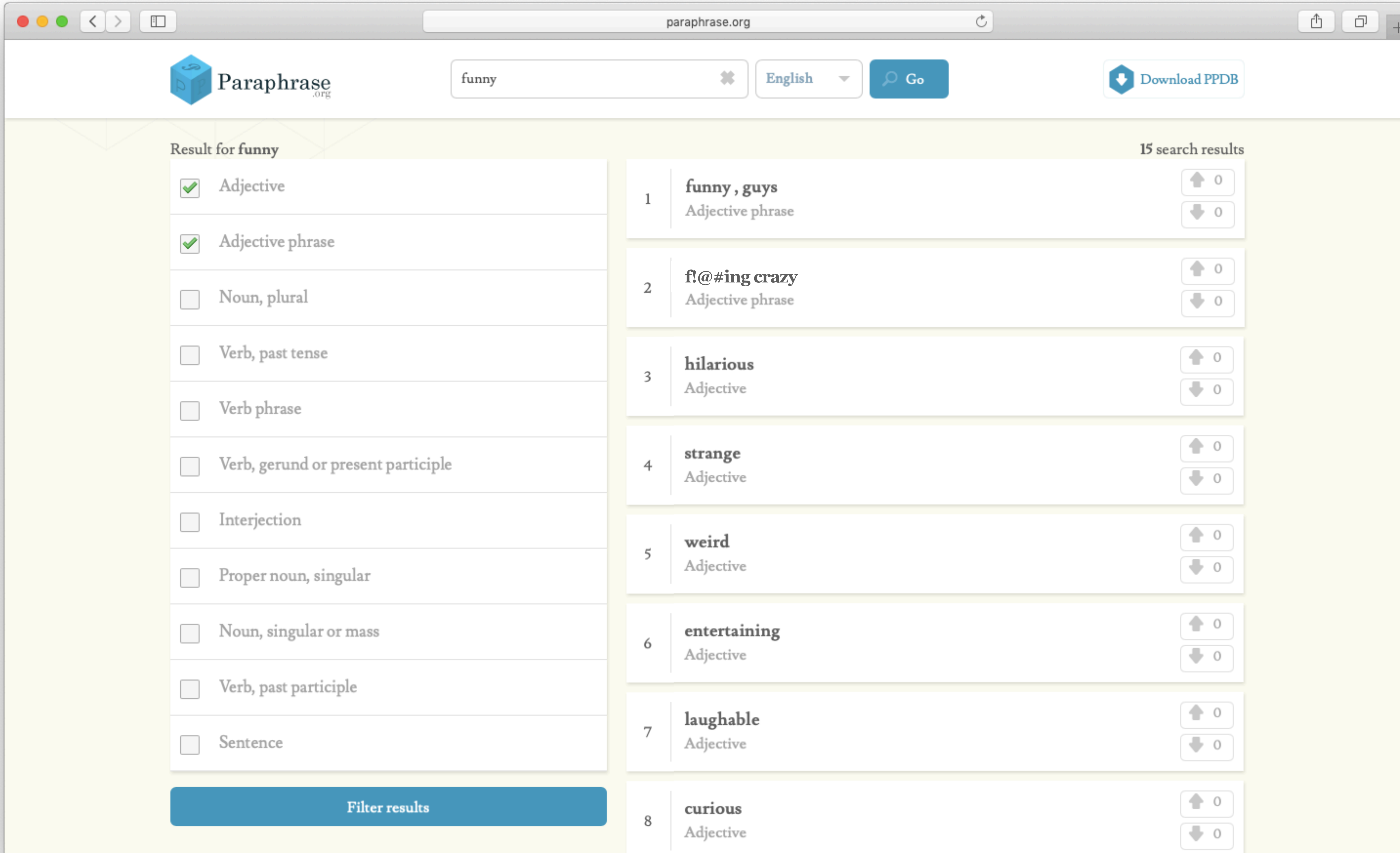
Paraphrasing with Bilingual Parallel Corpora. Colin Bannard and Chris Callison-Burch. ACL 2005.

PPDB 2.0

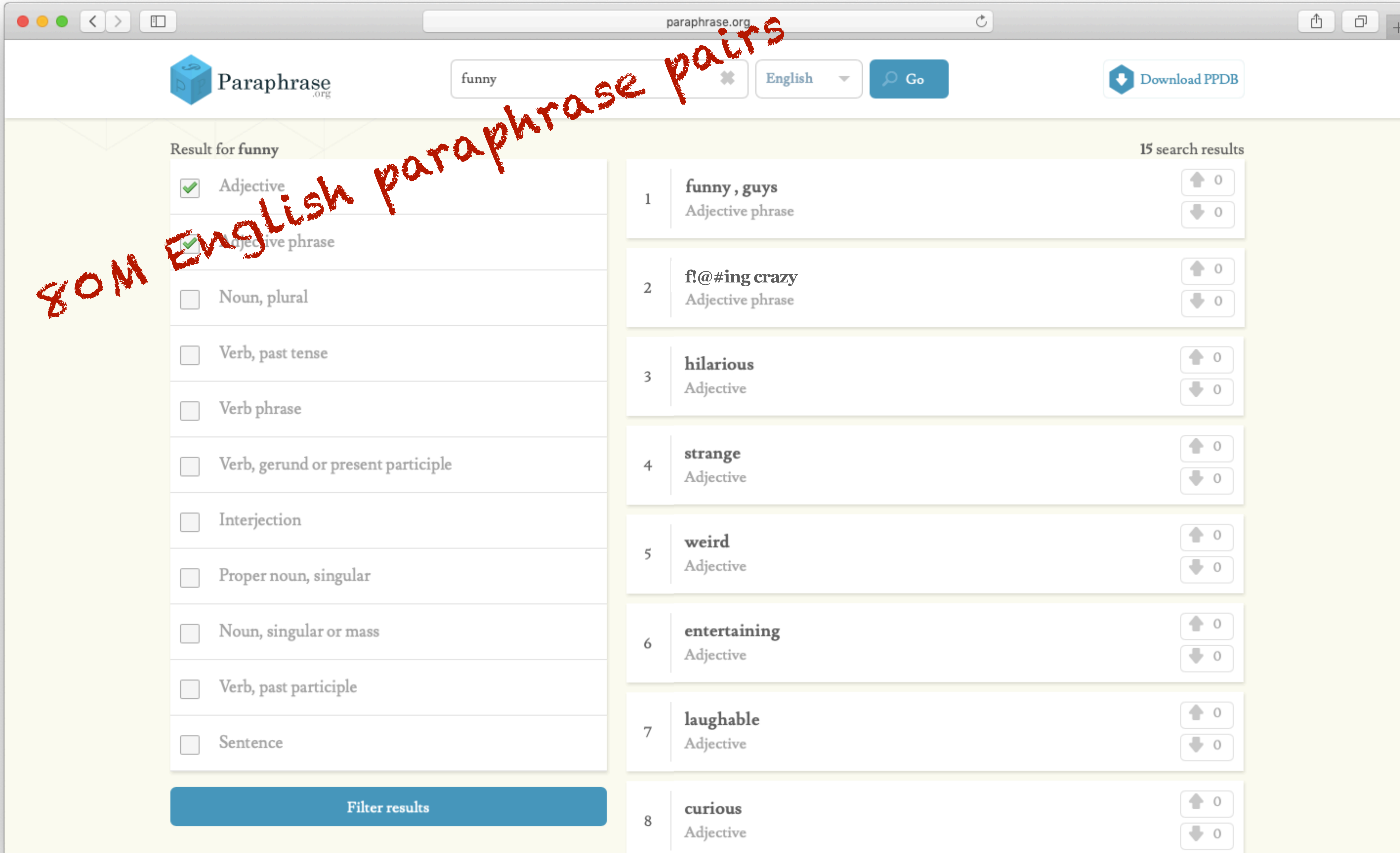
Re-ranked paraphrases better correlate with human judgments



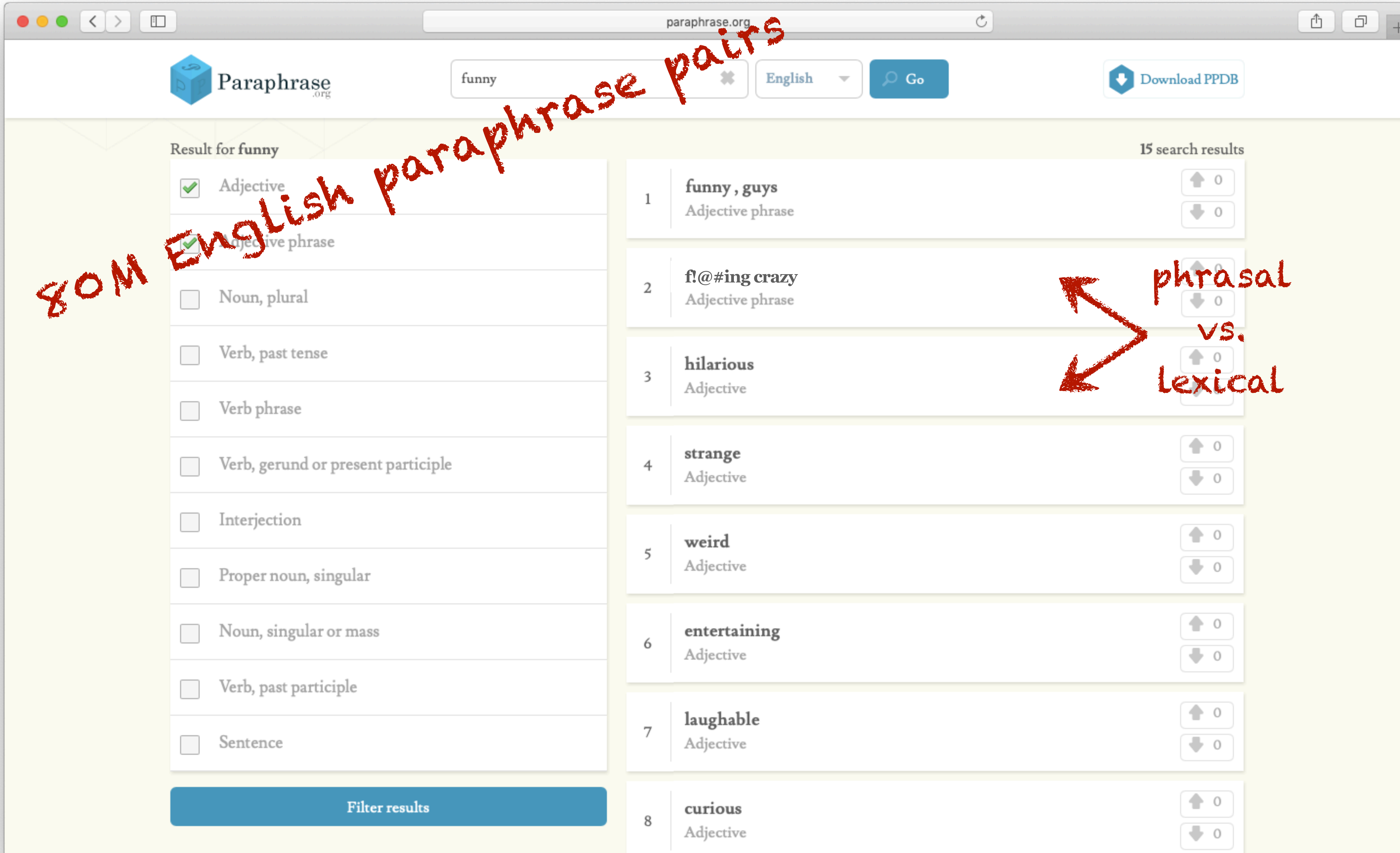
PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevich, Ben Van Durme, Chris Callison-Burch. ACL-2015.



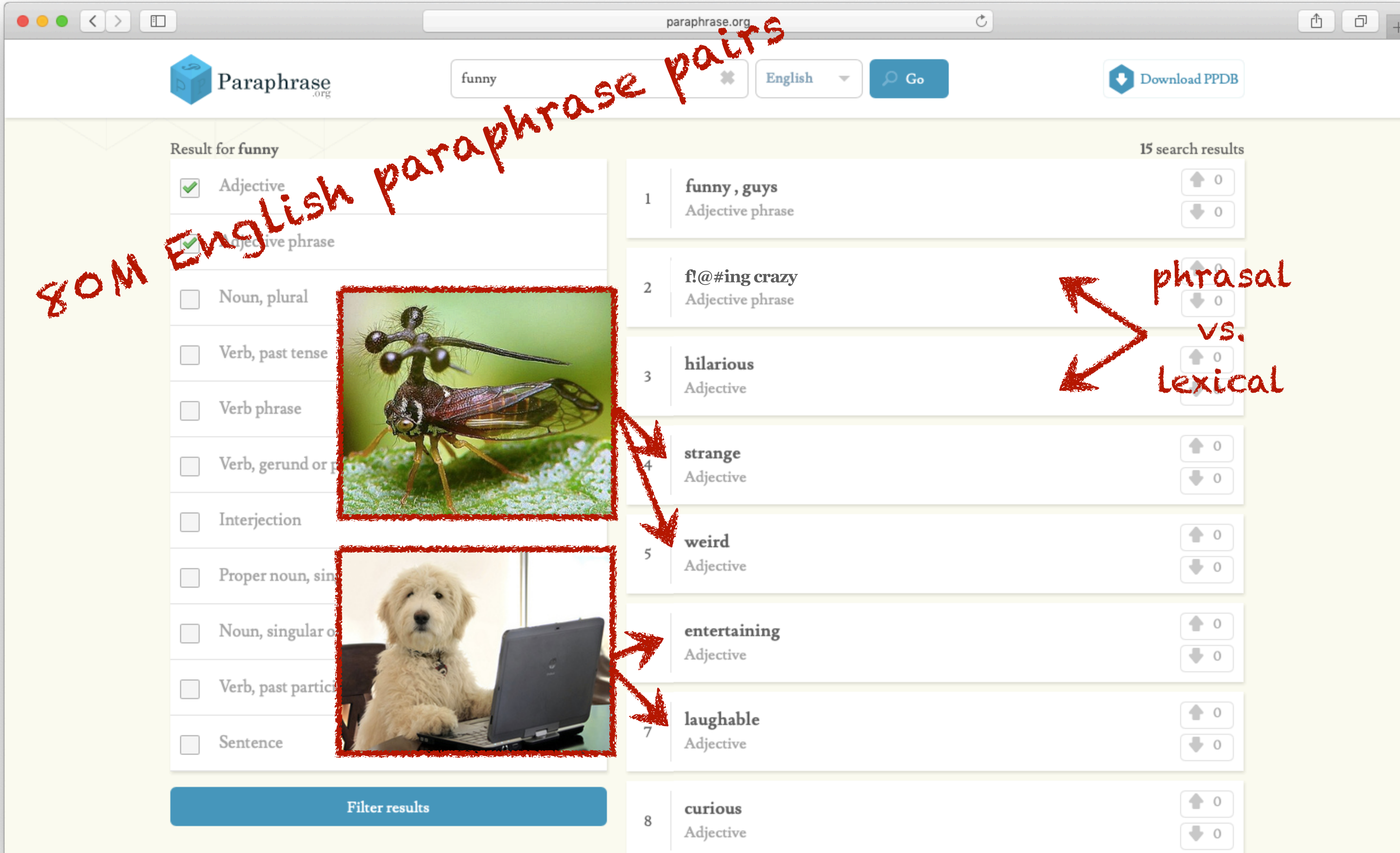
PPDB: The Paraphrase Database. Ganitkevitch et al., NAACL-HLT 2013.



PPDB: The Paraphrase Database. Ganitkevitch et al., NAACL-HLT 2013.



PPDB: The Paraphrase Database. Ganitkevitch et al., NAACL-HLT 2013.



PPDB: The Paraphrase Database. Ganitkevitch et al., NAACL-HLT 2013.

This Thesis

This Thesis

- Bilingually-induced paraphrases provide useful signal for modeling lexical semantics

This Thesis

- Bilingually-induced paraphrases provide useful signal for modeling lexical semantics

- Paraphrase-based signals are complementary to semantic information derived from monolingual distributional and pattern-based methods because:

This Thesis

- Bilingually-induced paraphrases provide useful signal for modeling lexical semantics
- Paraphrase-based signals are complementary to semantic information derived from monolingual distributional and pattern-based methods because:
  - the paraphrases of a word cover its multiple meanings,

This Thesis

- Bilingually-induced paraphrases provide useful signal for modeling lexical semantics
- Paraphrase-based signals are complementary to semantic information derived from monolingual distributional and pattern-based methods because:
    - the paraphrases of a word cover its multiple meanings,
    - paraphrases enable direct analysis of compositional phrases and their single-word equivalents,

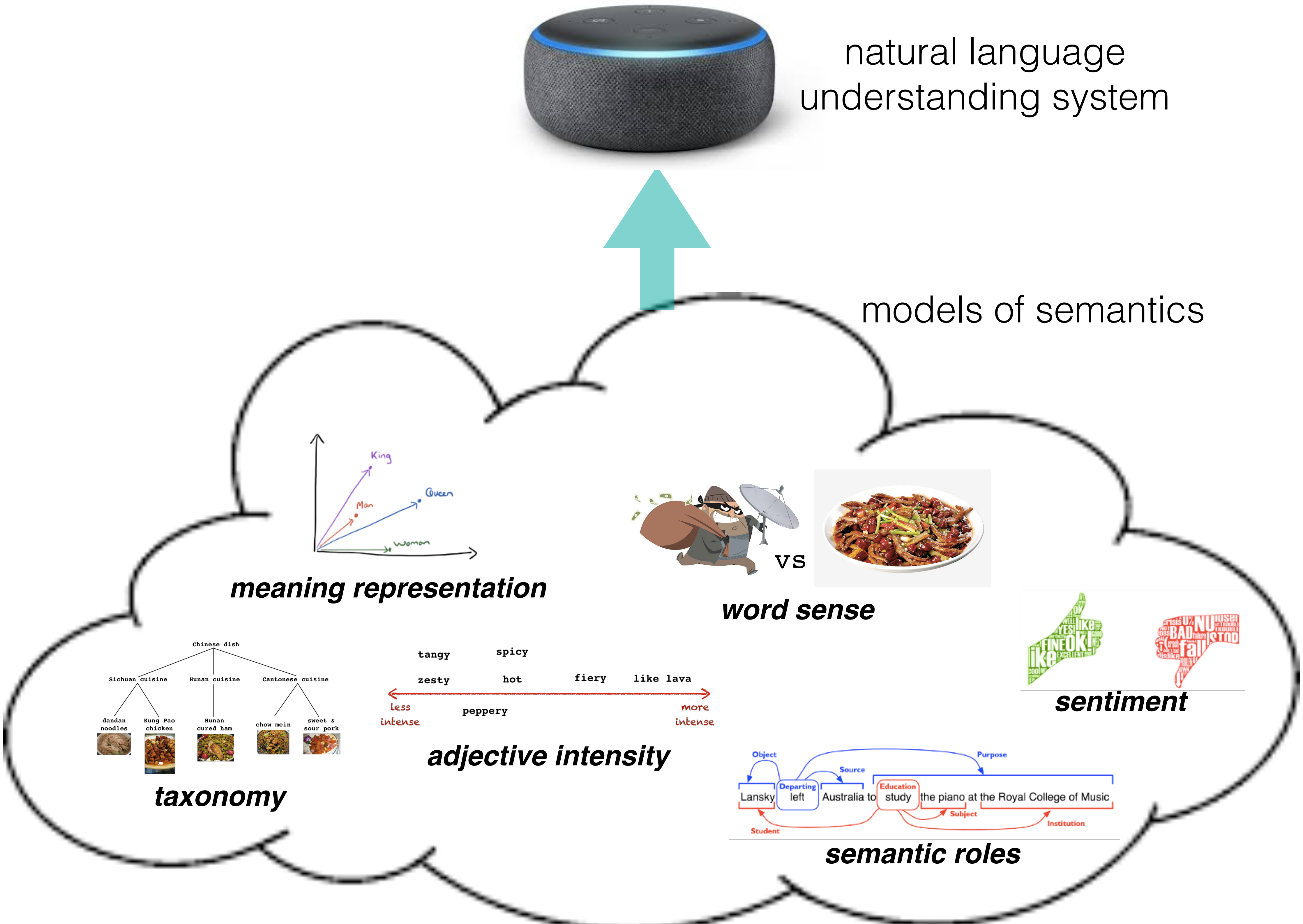
This Thesis

- Bilingually-induced paraphrases provide useful signal for modeling lexical semantics
- Paraphrase-based signals are complementary to semantic information derived from monolingual distributional and pattern-based methods because:
  - the paraphrases of a word cover its multiple meanings,
  - paraphrases enable direct analysis of compositional phrases and their single-word equivalents,
  - and paraphrases can be generated at scale.

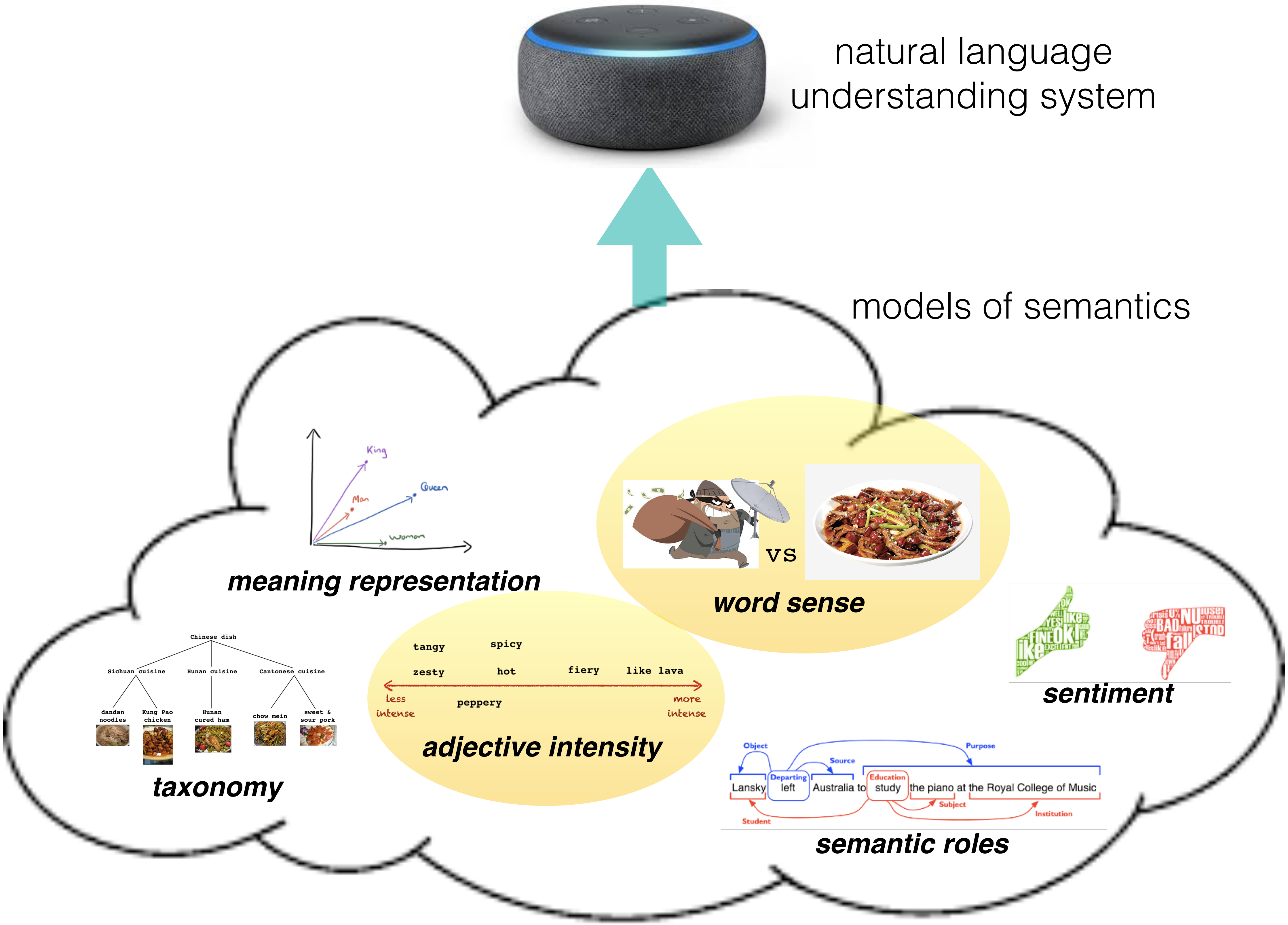
Putting this work into context

natural language
understanding system

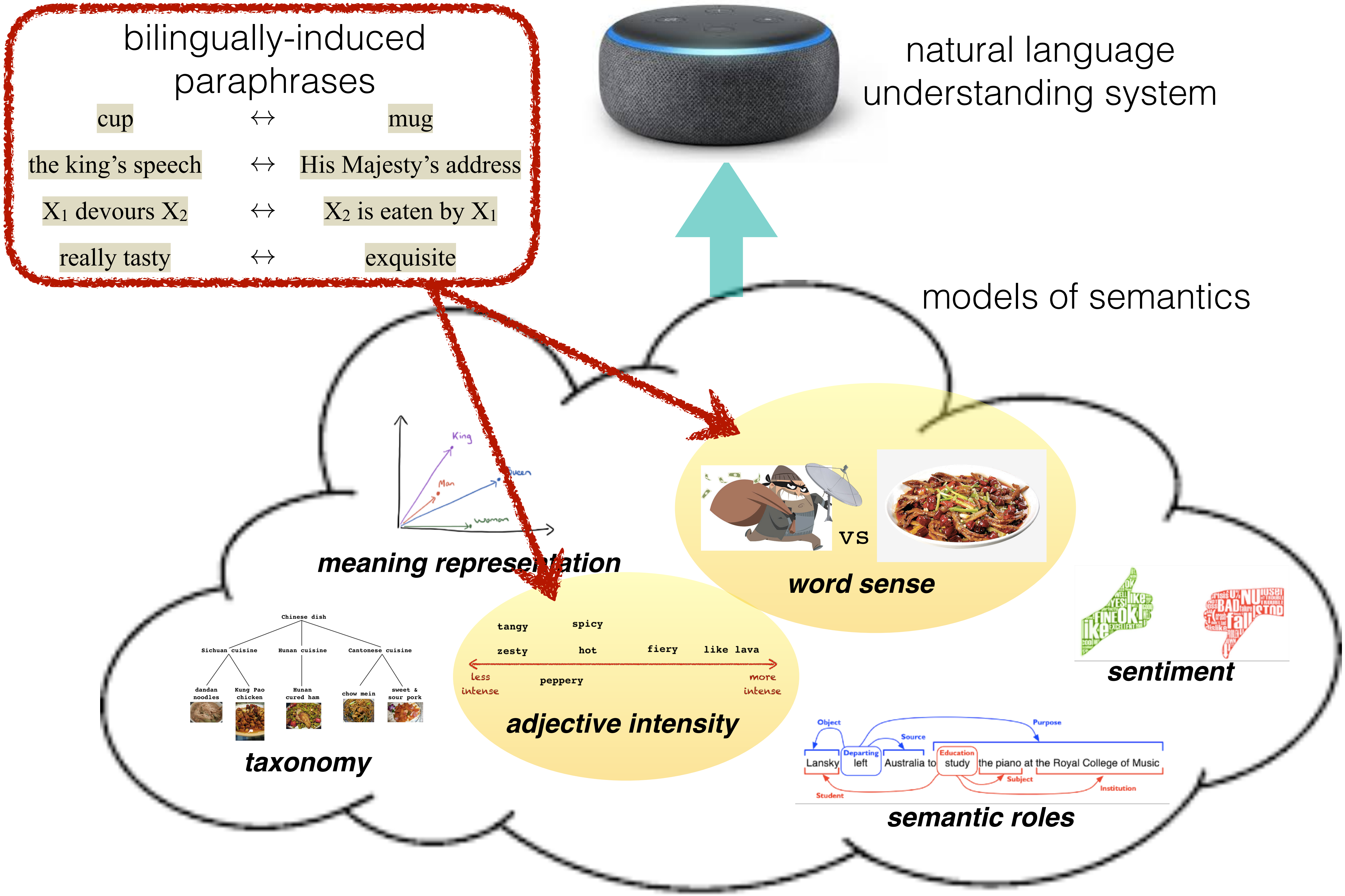
Putting this work into context



Putting this work into context



Putting this work into context



Motivation

Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

Learning Scalar Adjective Intensity

EMNLP 2018

Meaning-specific Examples of Word Use

In submission

Conclusion

Motivation

Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

Learning Scalar Adjective Intensity

EMNLP 2018

Meaning-specific Examples of Word Use

In submission

Conclusion

Motivation

Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

Learning Scalar Adjective Intensity

EMNLP 2018

Meaning-specific Examples of Word Use

In submission

Conclusion

Motivation

Using Paraphrases to Model Word Sense

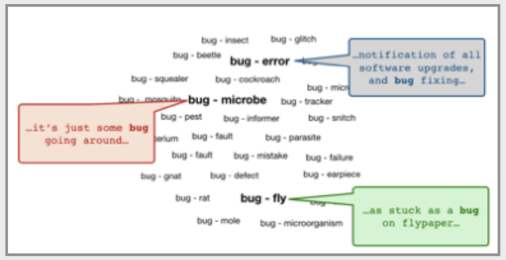
NAACL 2016; SENSE@EACL 2017

Learning Scalar Adjective Intensity

EMNLP 2018

Meaning-specific Examples of Word Use

In submission



Conclusion

Motivation

Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

Learning Scalar Adjective Intensity

EMNLP 2018

Meaning-specific Examples of Word Use

In submission

Conclusion

Motivation

Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

Learning Scalar Adjective Intensity

EMNLP 2018

Meaning-specific Examples of Word Use

In submission

Conclusion

"What's a Chinese **dish** that's not so **hot**?"

hot dish?

Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

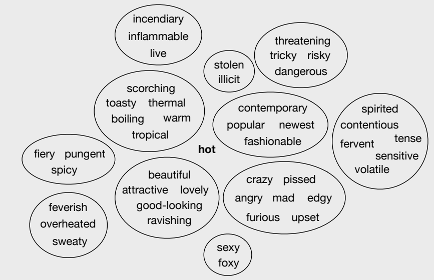
Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

- Claims:
    - Paraphrases can be used to model the different meanings of a target word through *sense clustering*

Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

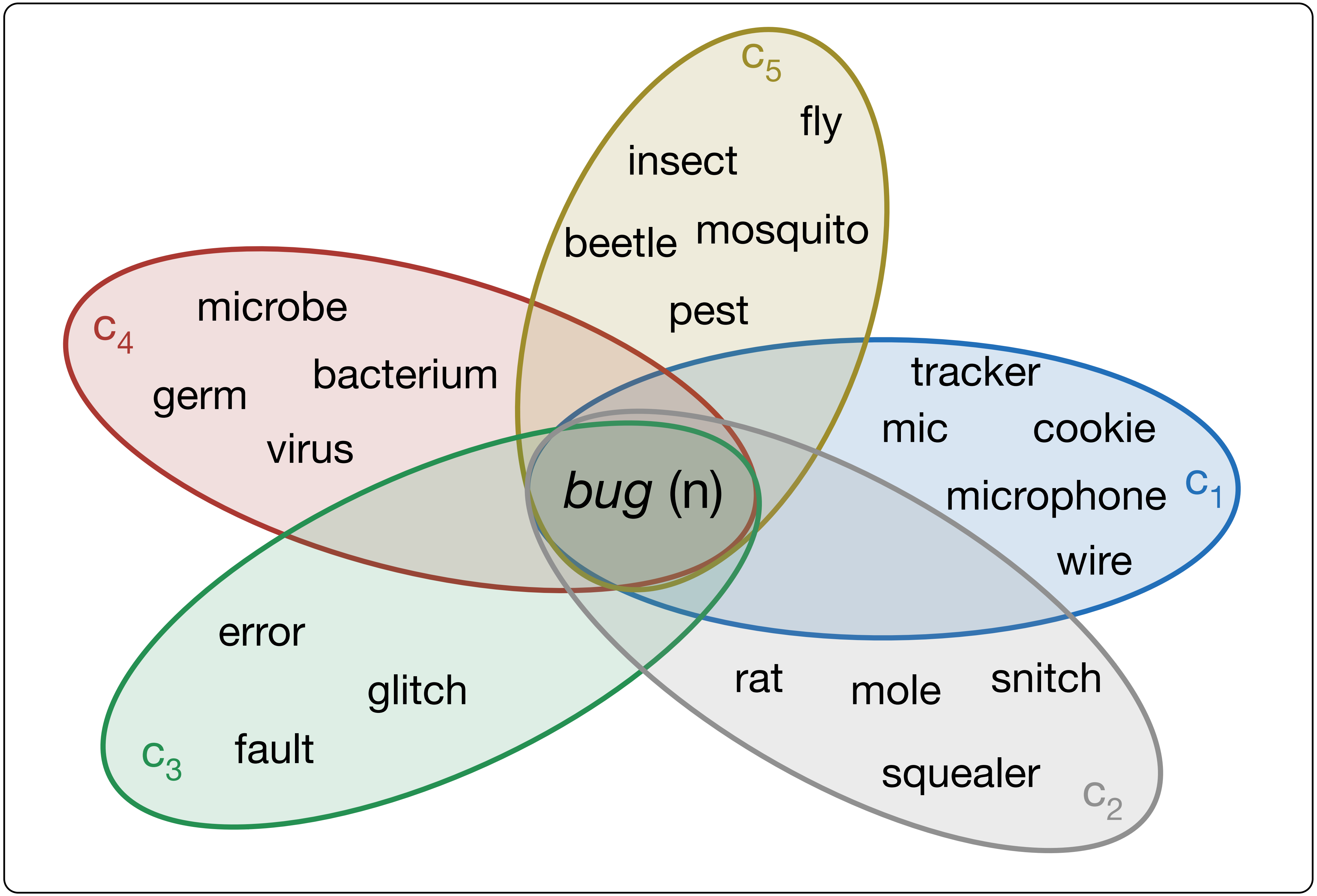


- Claims:
  - Paraphrases can be used to model the different meanings of a target word through *sense clustering*
  - The resulting *sense clusters* can be used to help find the most applicable substitutes for a target word in context

Given a paraphrase set for a target word…

fly
insect
microbe
error
pest
bacterium
beetle
snitch
tracker
mole
cookie
bug (n)
mic
germ
microphone
virus
wire
rat
glitch
fault
mosquito
squealer

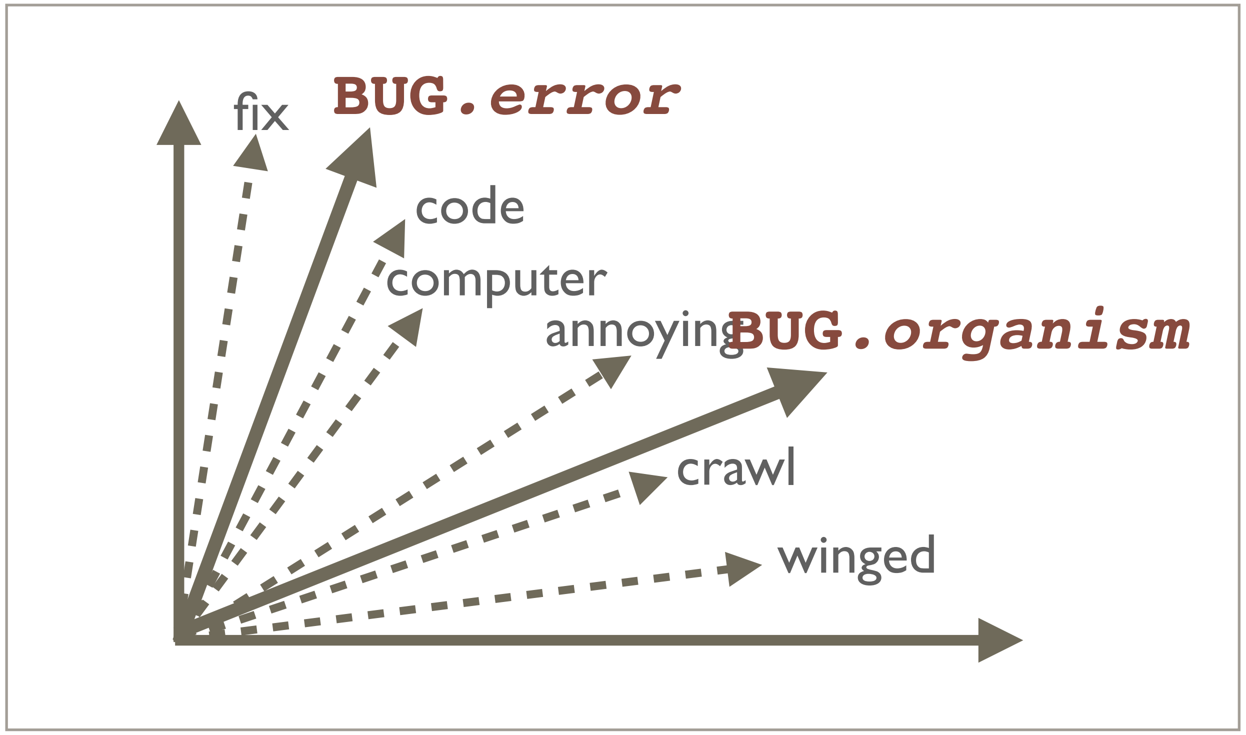
…we can model the different meanings of the target word by clustering its paraphrases.



This goal is closely related to earlier work on Word Sense Induction.

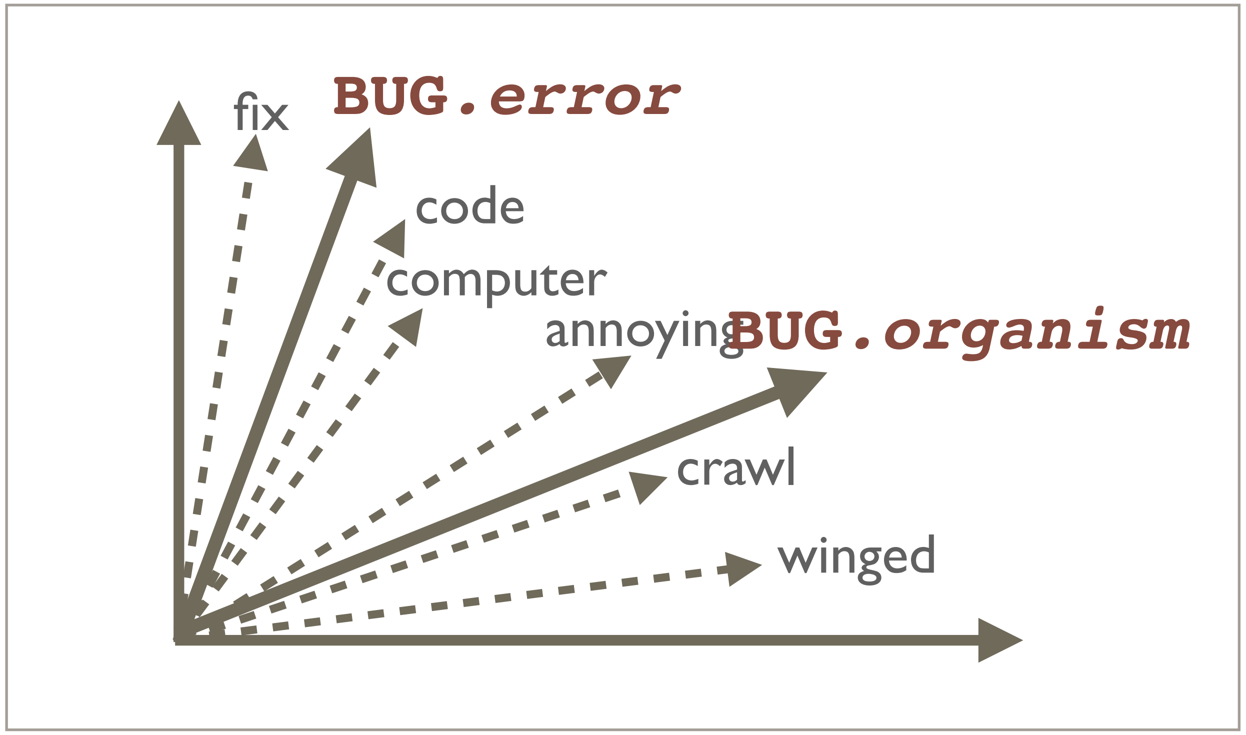
This goal is closely related to earlier work on Word Sense Induction.

- Clustering contexts or similar words in the same language
    - Schutze; Pantel & Lin; others

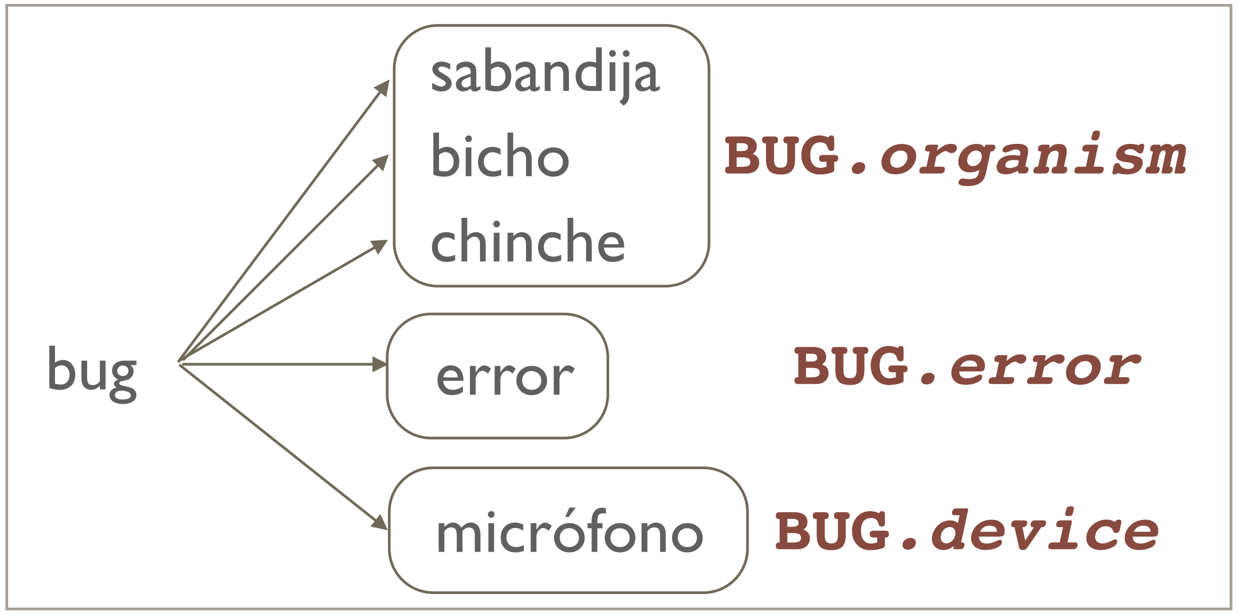


This goal is closely related to earlier work on Word Sense Induction.

- Clustering contexts or similar words in the same language
    - Schutze; Pantel & Lin; others

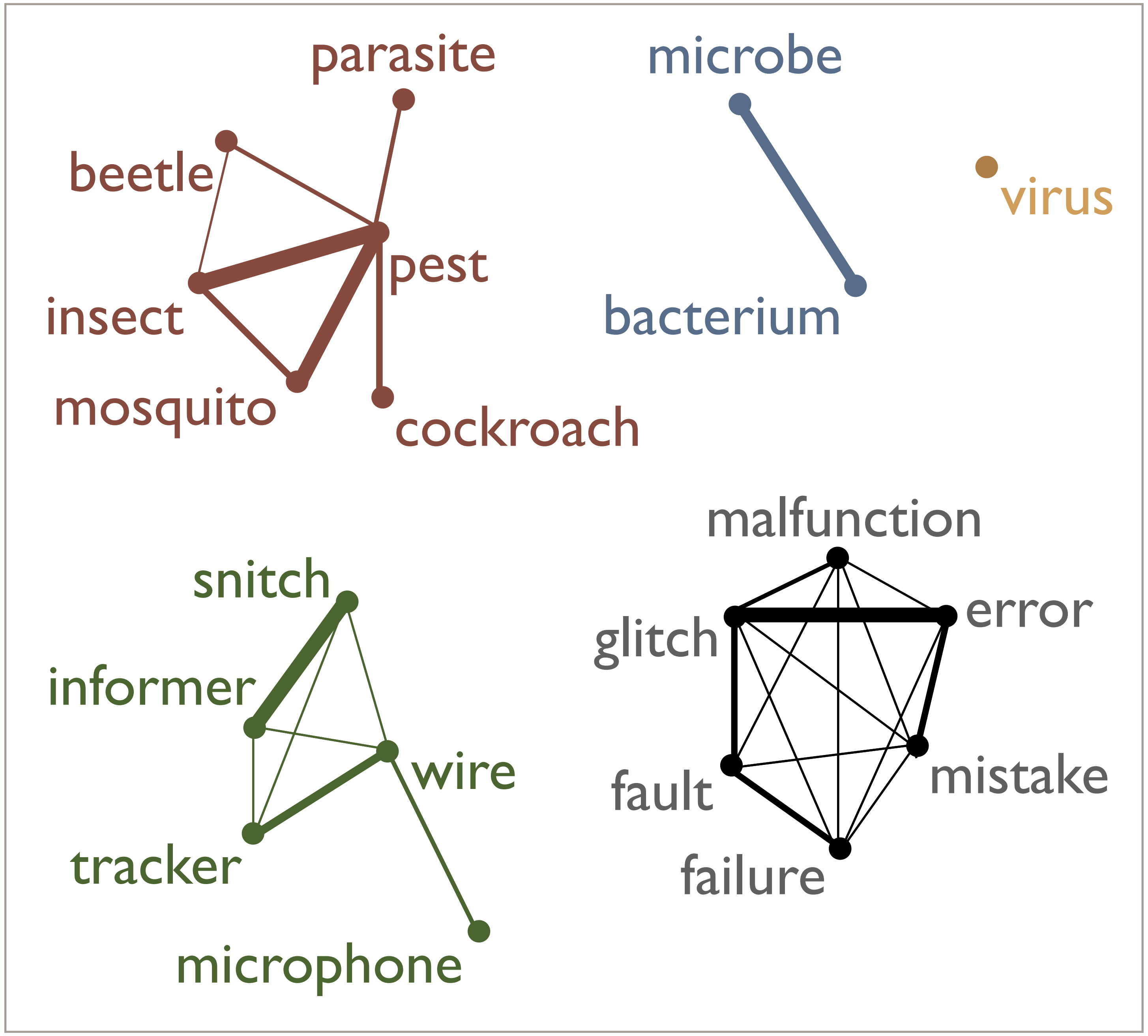


- Aligning senses to foreign translations
    - Gale et al.; Diab & Resnik; Apidianaki; others



This goal is closely related to earlier work on Word Sense Induction.

- Semantic paraphrase clustering (SEMCLUST) (Apidianaki et al. 2014)
    - Demonstrated that sense distinctions exist in PPDB
    - We use this method as a baseline

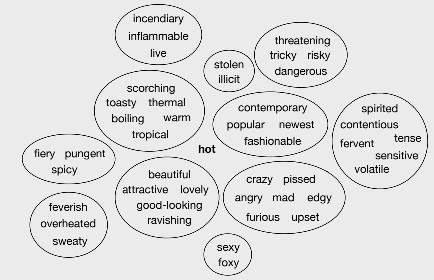


Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

Using Paraphrases to Model Word Sense

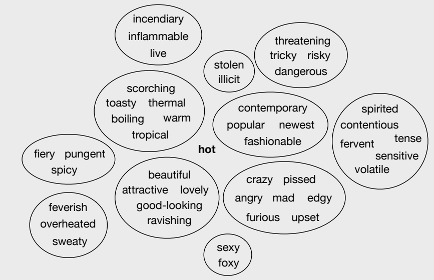
NAACL 2016; SENSE@EACL 2017



- Goals:

Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017



- Goals:
  - Validate that paraphrases can be clustered to model different word meanings

Using Paraphrases to Model Word Sense

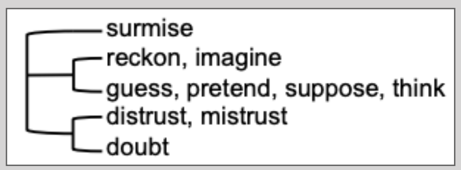
NAACL 2016; SENSE@EACL 2017

- Goals:
  - Validate that paraphrases can be clustered to model different word meanings
  - Compare paraphrase-based semantic similarity metrics with other signal types for clustering

Our experiments cluster a target word's paraphrase set to model its different meanings.

<u>2 clustering algorithms</u>

- HGFC



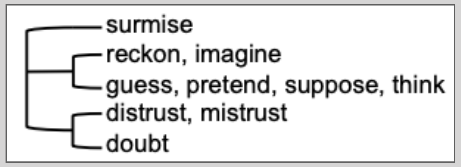
- Spectral

k=3
c_1: reckon, pretend, think, imagine
c_2: guess, suppose, surmise
c_3: distrust, doubt, mistrust

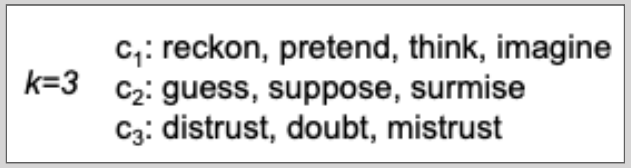
Our experiments cluster a target word's paraphrase set to model its different meanings.

2 clustering algorithms

- HGFC



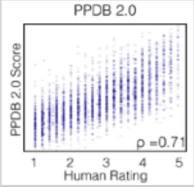
- Spectral



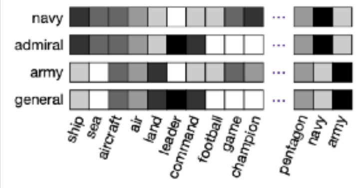
x

5 similarity metrics

- PPDBScore (direct)



- PPDBScore (2nd-order) (two types)
- Distributional Similarity

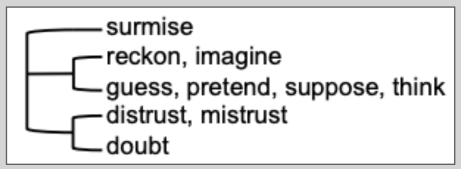


- Foreign Translation Similarity

Our experiments cluster a target word's paraphrase set to model its different meanings.

2 clustering algorithms

- HGFC



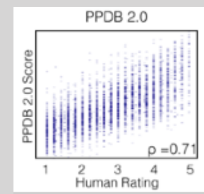
- Spectral

k=3	c_1: reckon, pretend, think, imagine
	c_2: guess, suppose, surmise
	c_3: distrust, doubt, mistrust

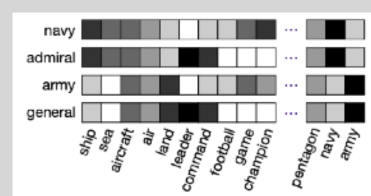
x

5 similarity metrics

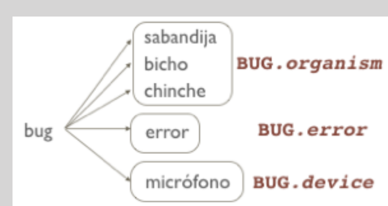
- PPDBScore (direct)



- PPDBScore (2nd-order) (two types)
- Distributional Similarity



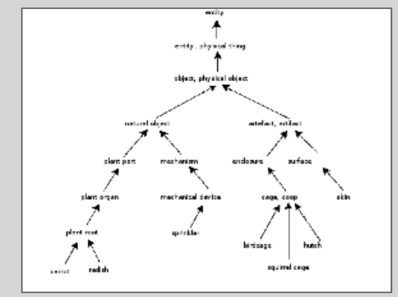
- Foreign Translation Similarity



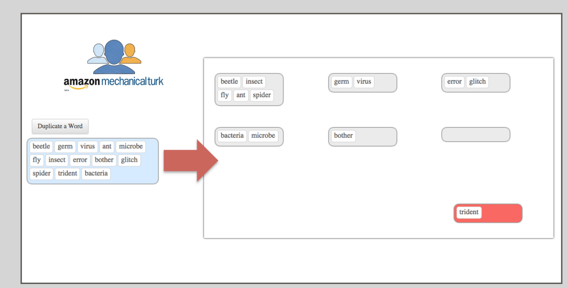
x

2 human-generated sense inventories

- WordNet+



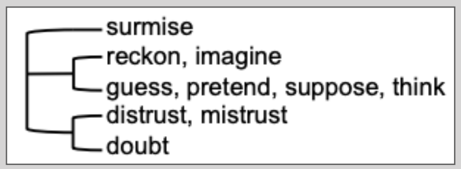
- CrowdClusters



Our experiments cluster a target word's paraphrase set to model its different meanings.

2 clustering algorithms

- HGFC



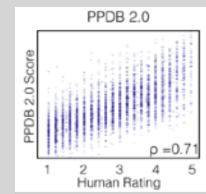
- Spectral

k=3
c_1: reckon, pretend, think, imagine
c_2: guess, suppose, surmise
c_3: distrust, doubt, mistrust

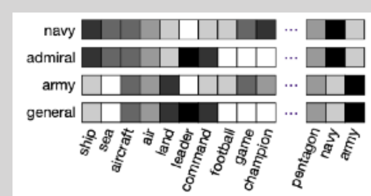
x

5 similarity metrics

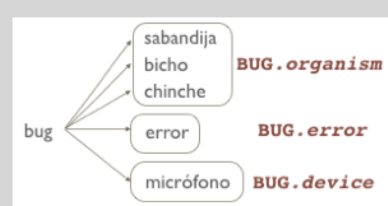
- PPDBScore (direct)



- PPDBScore (2nd-order) (two types)
- Distributional Similarity



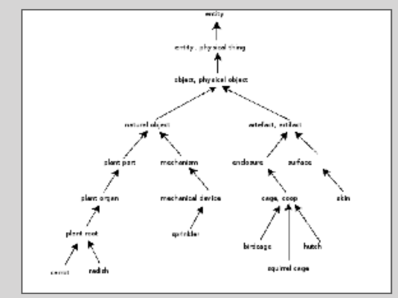
- Foreign Translation Similarity



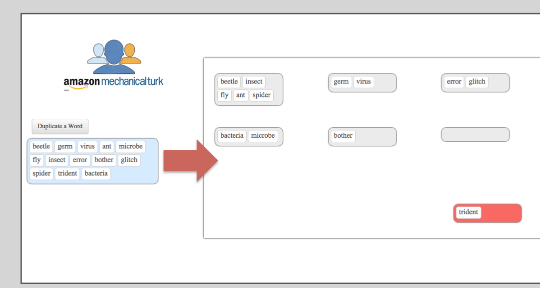
x

2 human-generated sense inventories

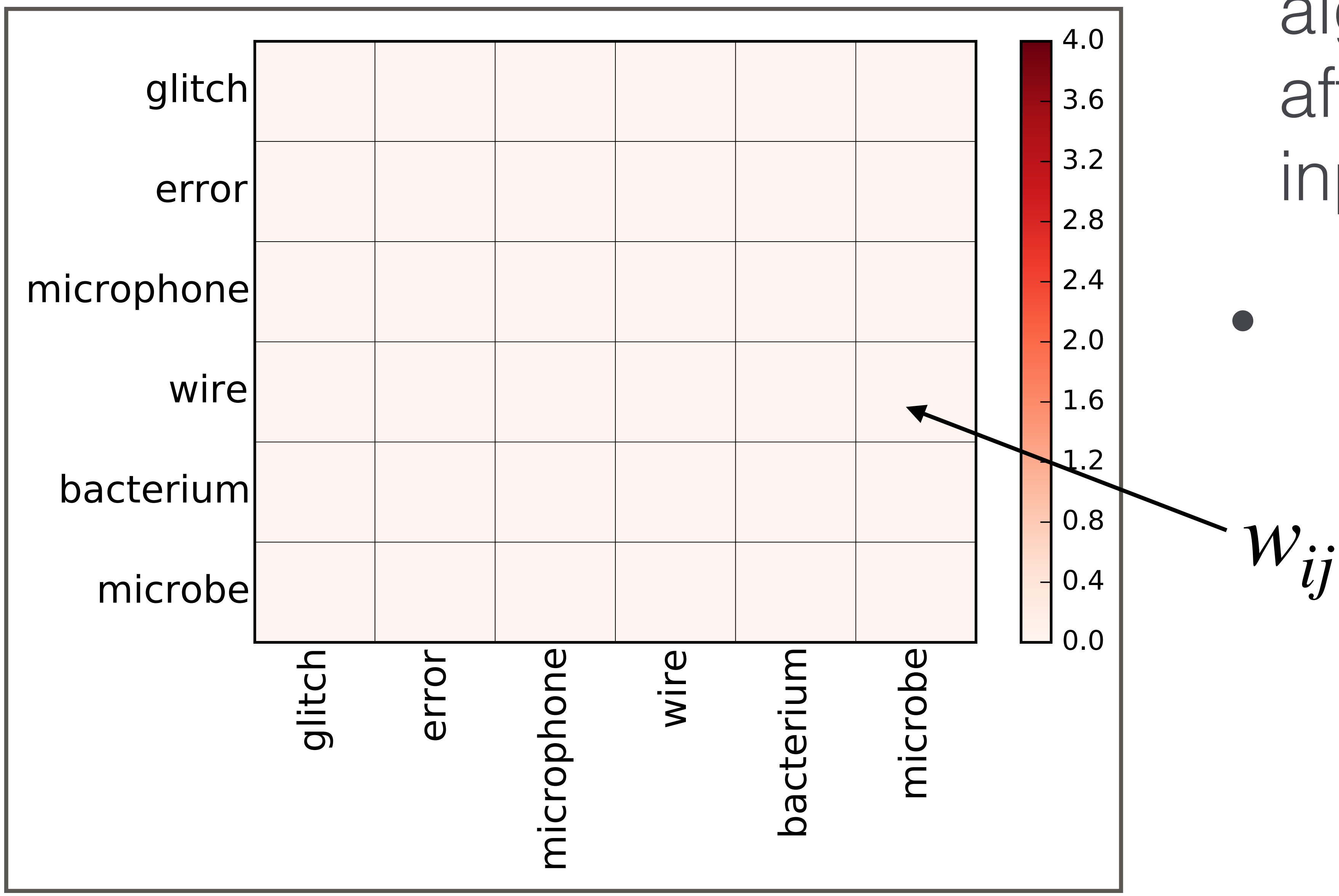
- WordNet+



- CrowdClusters

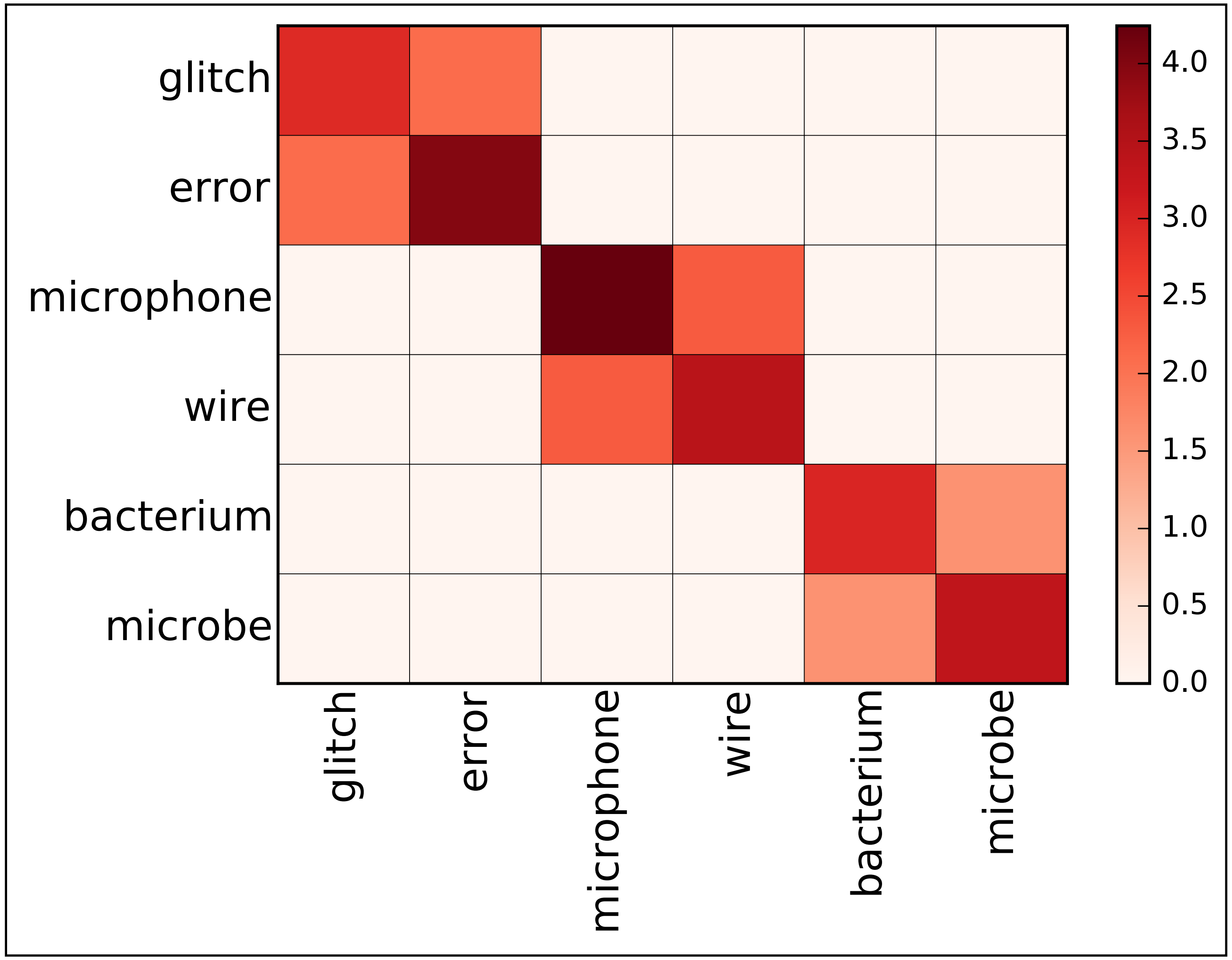


We vary **similarity metrics**

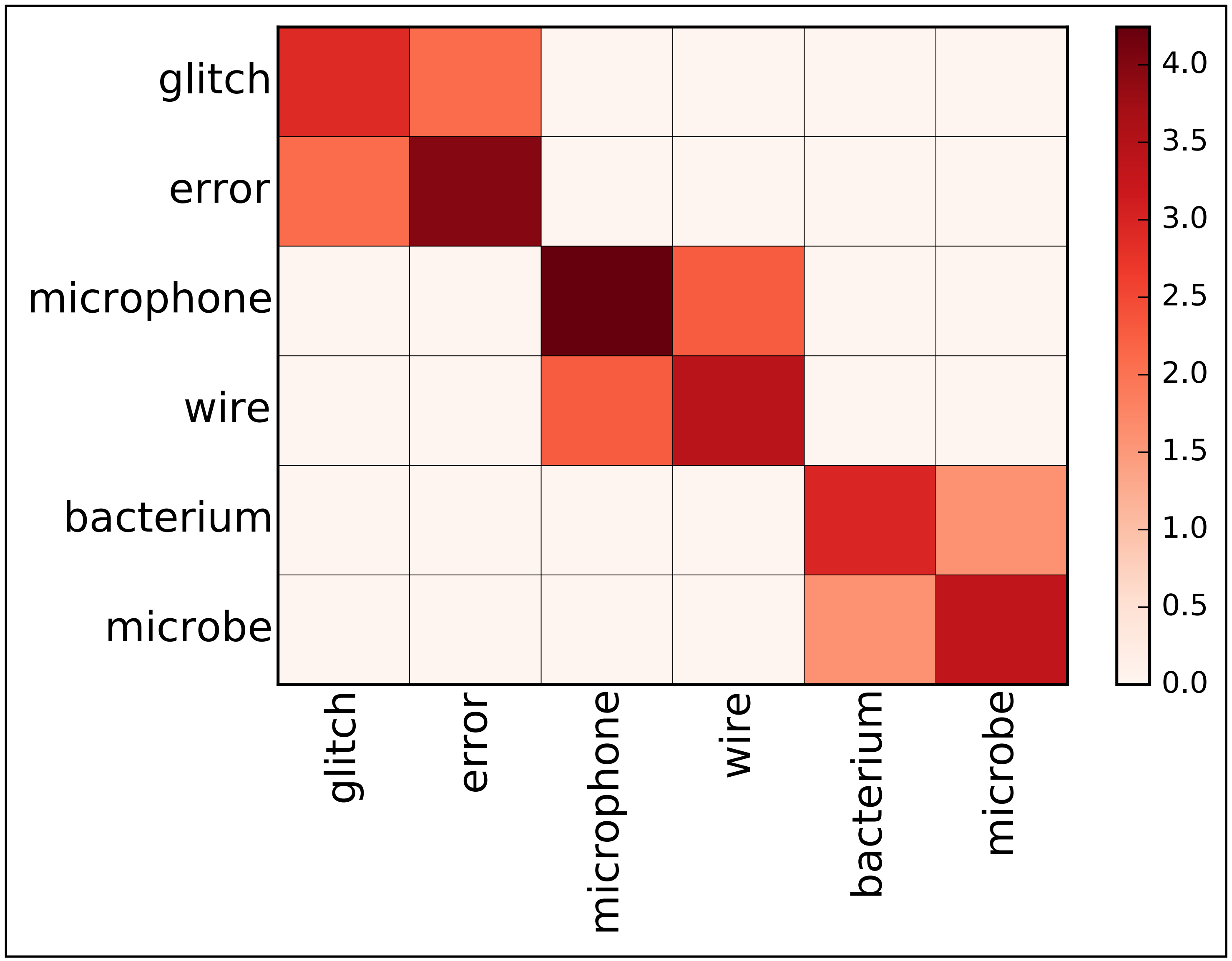


- Our clustering algorithm takes an affinity matrix as input
    - How should we fill it?

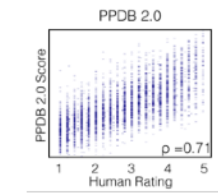
We vary **similarity metrics**



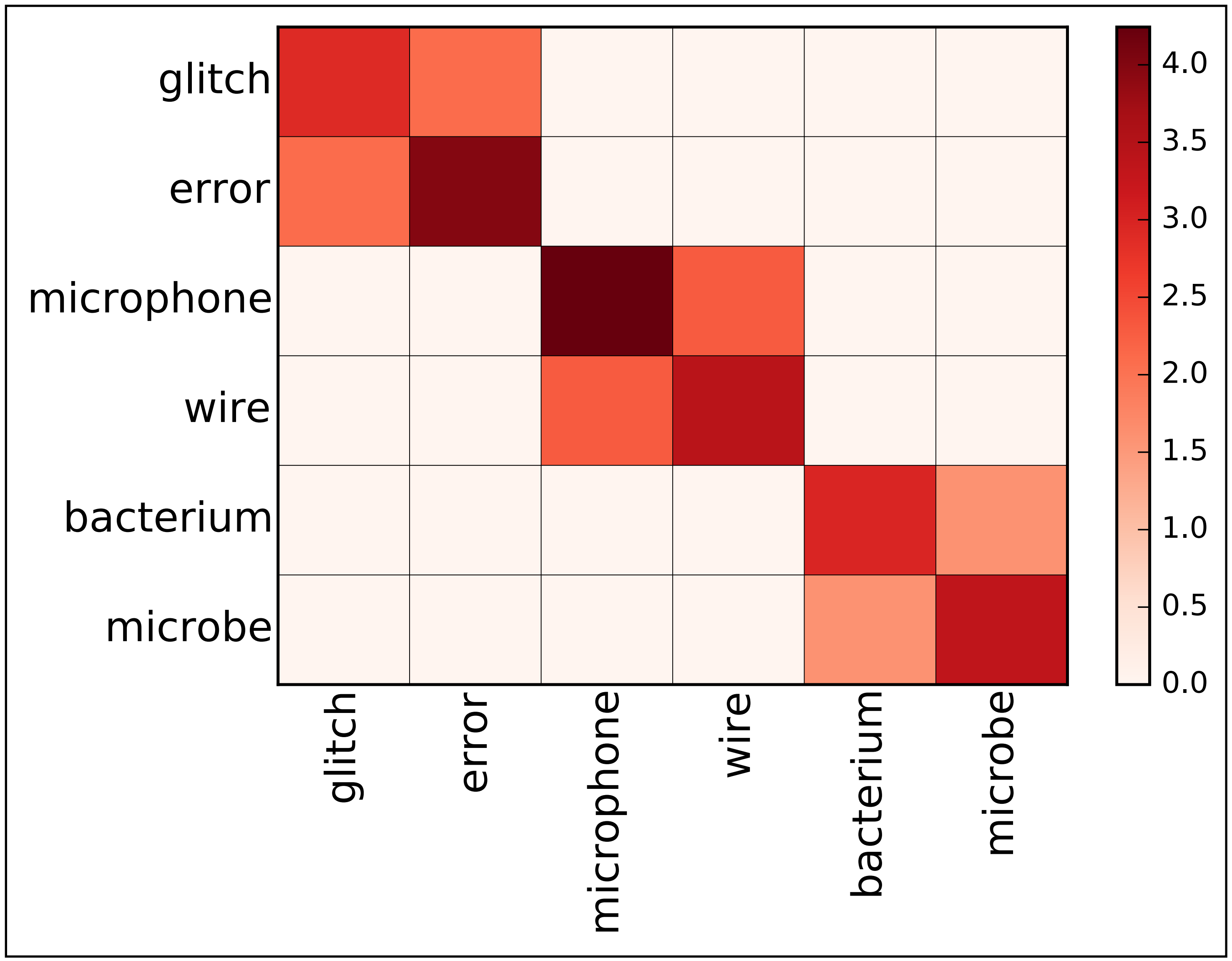
We vary **similarity metrics**



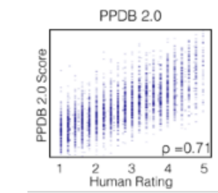
- Direct PPDB Score ($\text{sim}_{\text{PPDB2.0}}$)



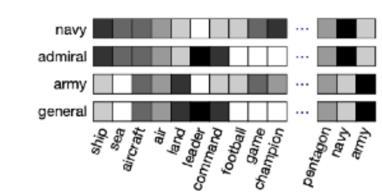
We vary **similarity metrics**



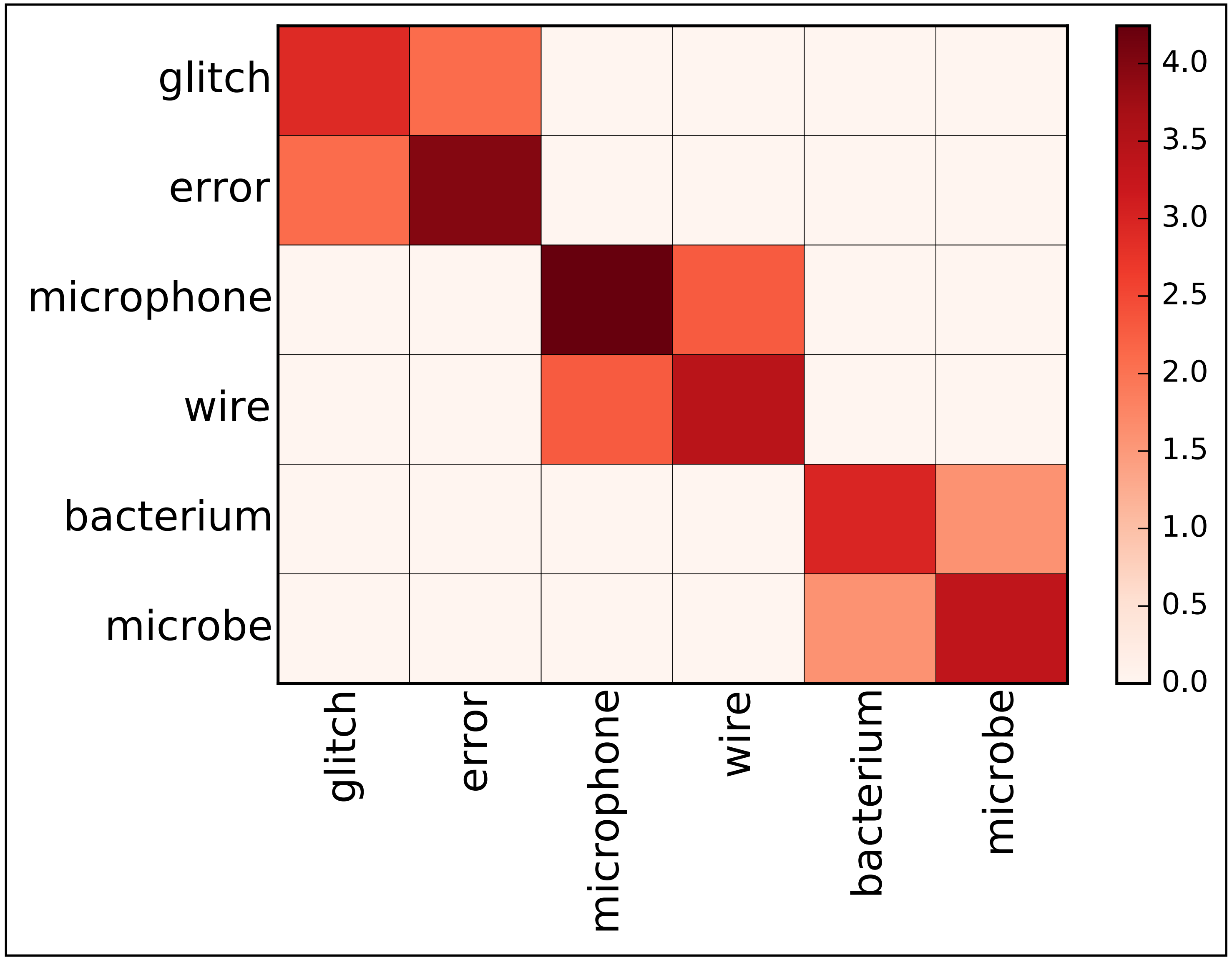
- Direct PPDB Score ($sim_{PPDB2.0}$)



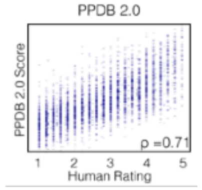
- Distributional Similarity ($sim_{DISTRIB}$)
  - cosine similarity of **`word2vec`** embeddings



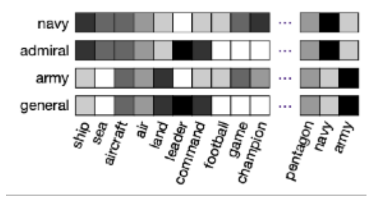
We vary **similarity metrics**



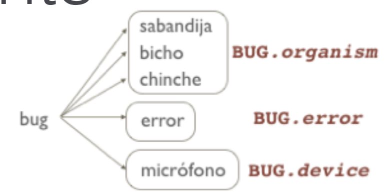
- Direct PPDB Score ($\text{sim}_{\text{PPDB2.0}}$)



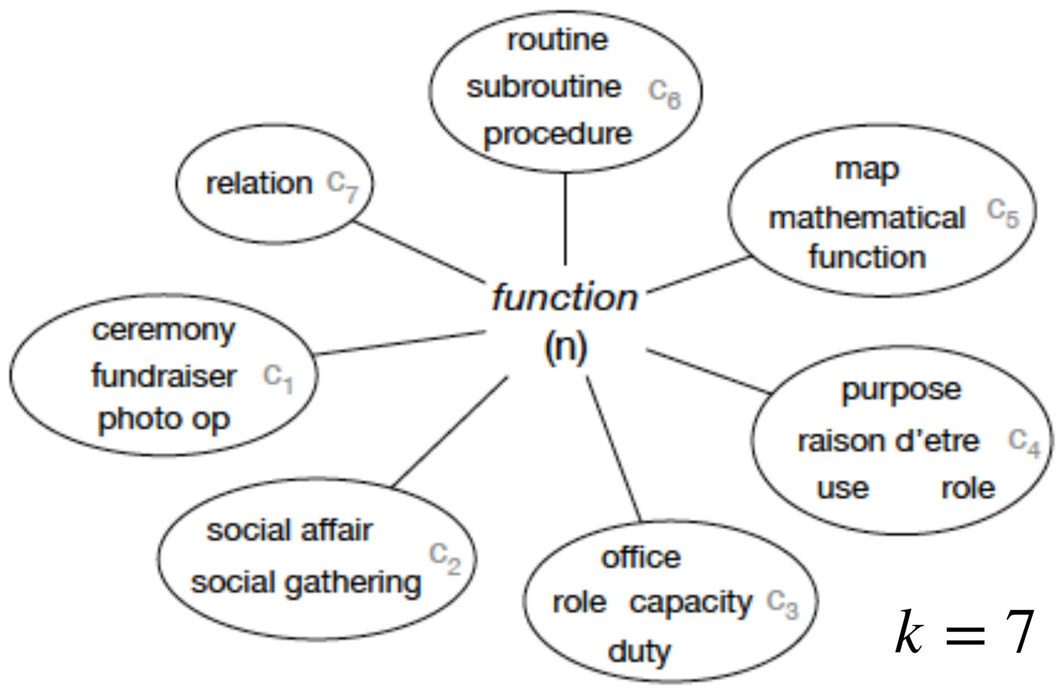
- Distributional Similarity ($\text{sim}_{\text{DISTRIB}}$)
  - cosine similarity of **`word2vec`** embeddings



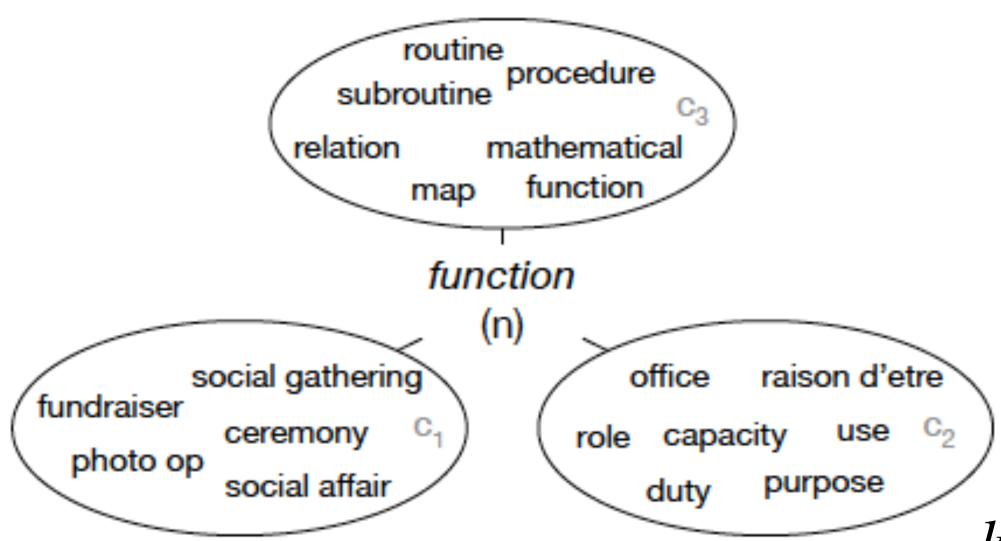
- Foreign Alignments ($\text{sim}_{\text{TRANS}}$)
  - cosine sim of ‘translation vectors’



The **number of senses** is chosen automatically



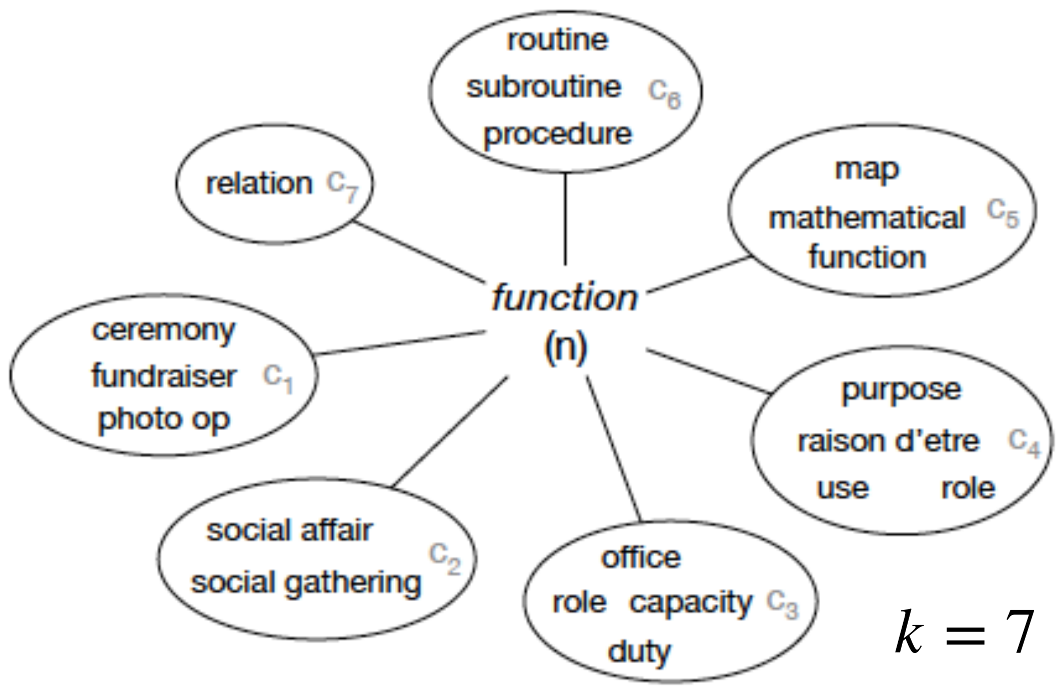
$k = 7$



$k = 3$

- Silhouette coefficient
    - Aims to find an 'optimal' number of clusters

The **number of senses** is chosen automatically



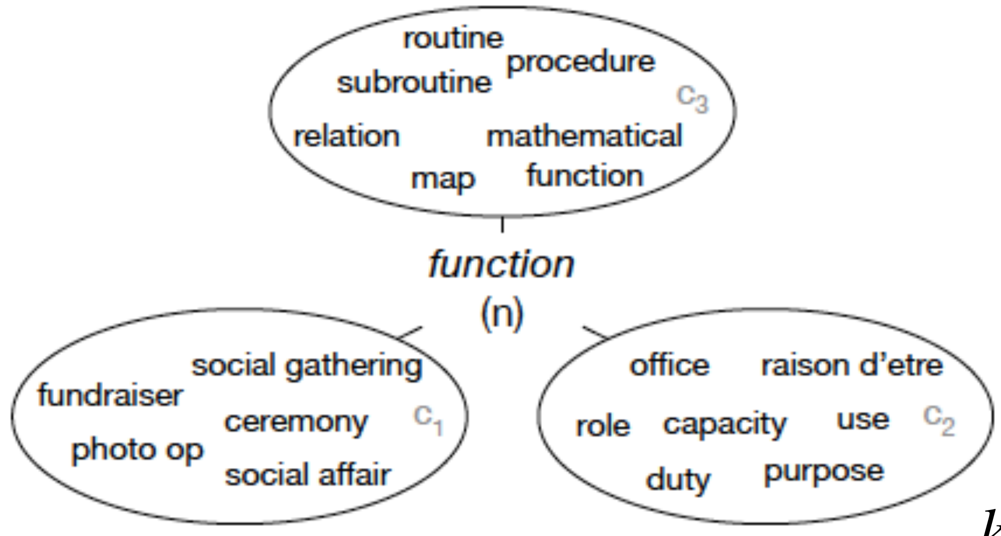
$k = 7$

- Silhouette coefficient
  - Aims to find an 'optimal' number of clusters

Given instance i

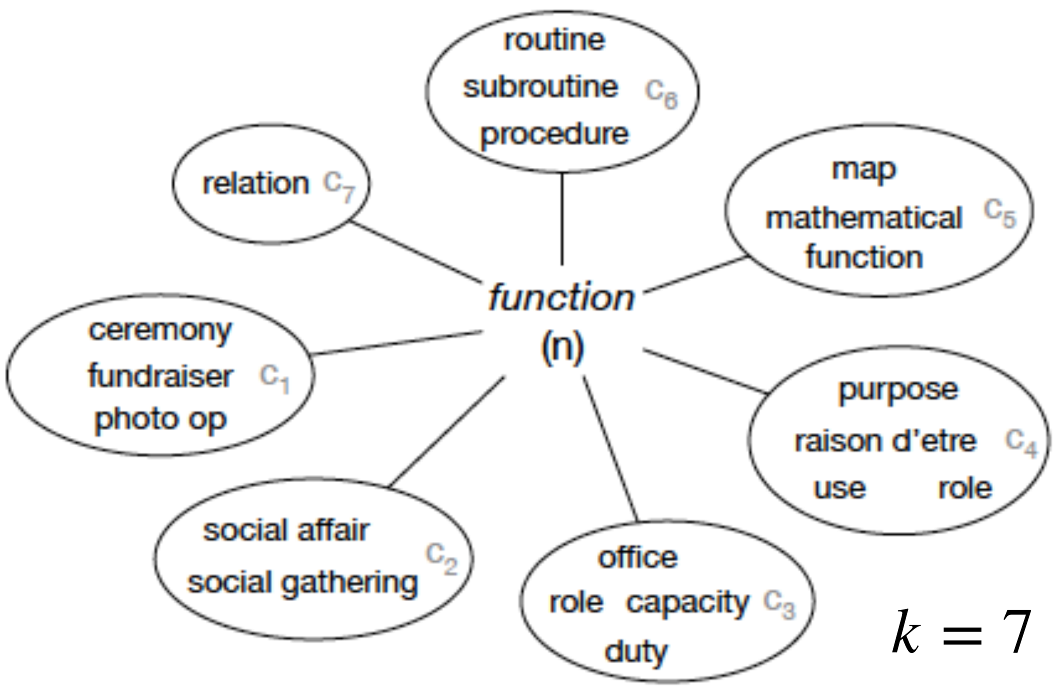
$a(i) =$ avg. same-cluster dist.

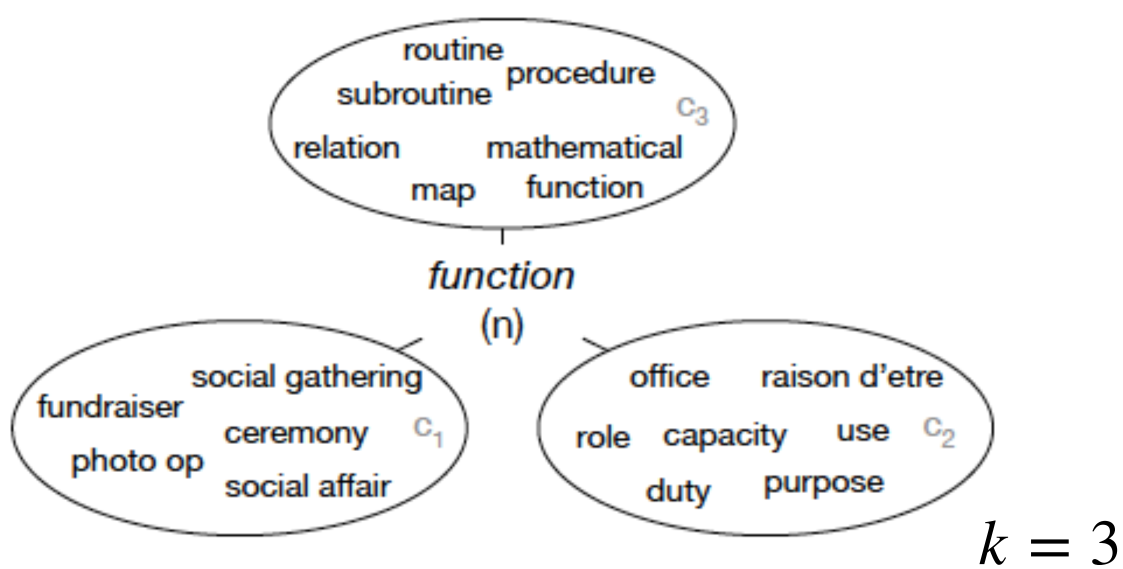
$b(i) =$ avg. nearest-cluster dist.



$k = 3$

The **number of senses** is chosen automatically





- Silhouette coefficient
  - Aims to find an 'optimal' number of clusters

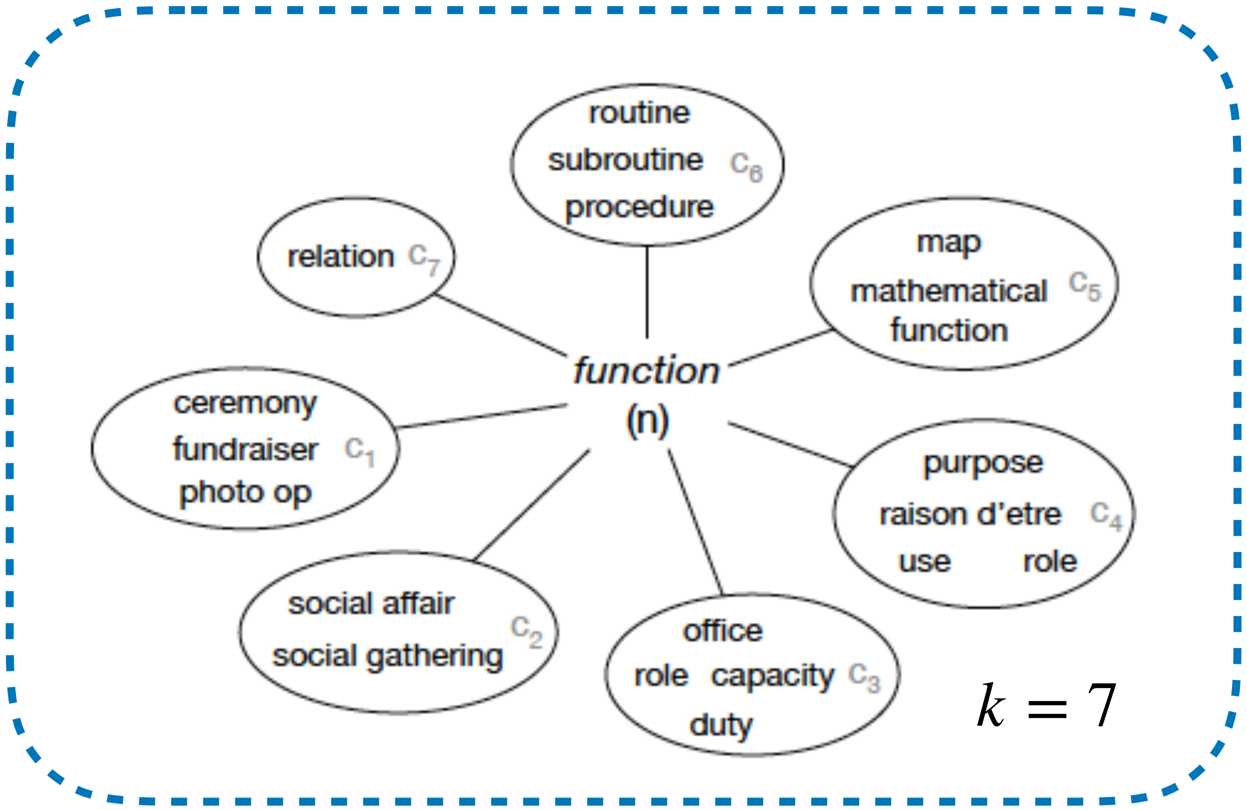
Given instance i

$$a(i) = \text{avg. same-cluster dist.}$$

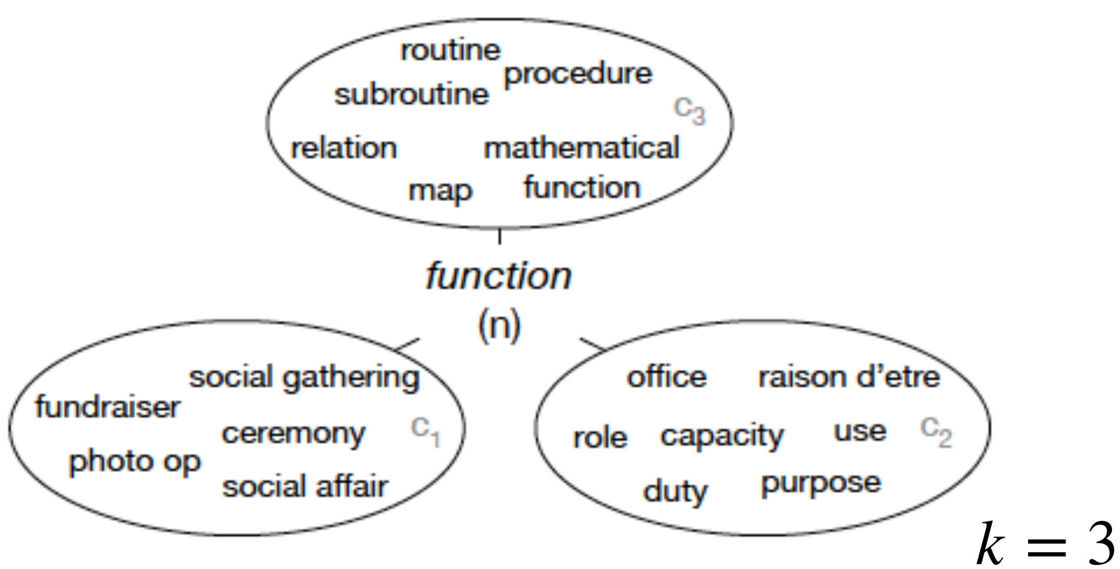
$$b(i) = \text{avg. nearest-cluster dist.}$$

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$$

The **number of senses** is chosen automatically



“optimal” clustering



- Silhouette coefficient
    - Aims to find an ‘optimal’ number of clusters

Given instance i

$$a(i) = \text{avg. same-cluster dist.}$$

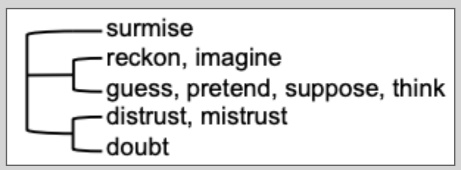
$$b(i) = \text{avg. nearest-cluster dist.}$$

$$s(i) = \frac{b(i) - a(i)}{\mathbf{max}\{b(i), a(i)\}}$$

Our experiments cluster a target word's paraphrase set to model its different meanings.

2 clustering algorithms

- HGFC



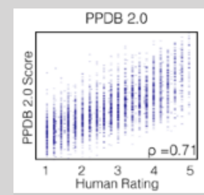
- Spectral

k=3
c_1: reckon, pretend, think, imagine
c_2: guess, suppose, surmise
c_3: distrust, doubt, mistrust

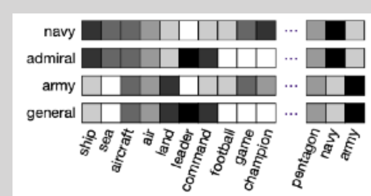
x

5 similarity metrics

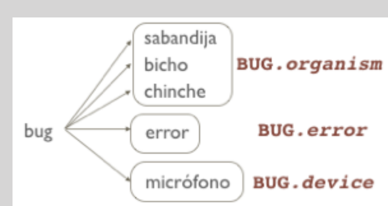
- PPDBScore (direct)



- PPDBScore (2[nd]-order) (two types)
- Distributional Similarity



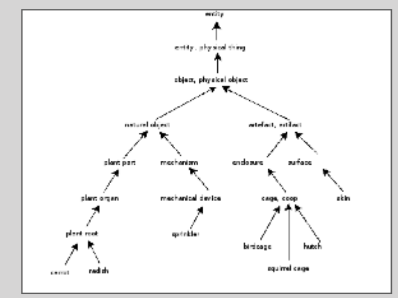
- Foreign Translation Similarity



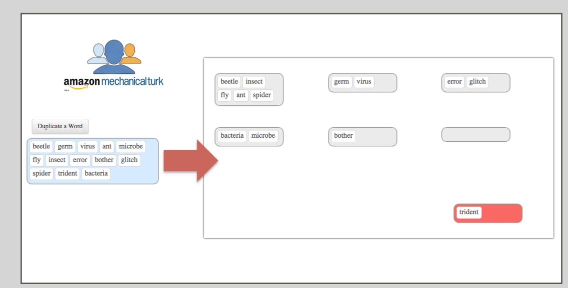
x

2 human-generated sense inventories

- WordNet+

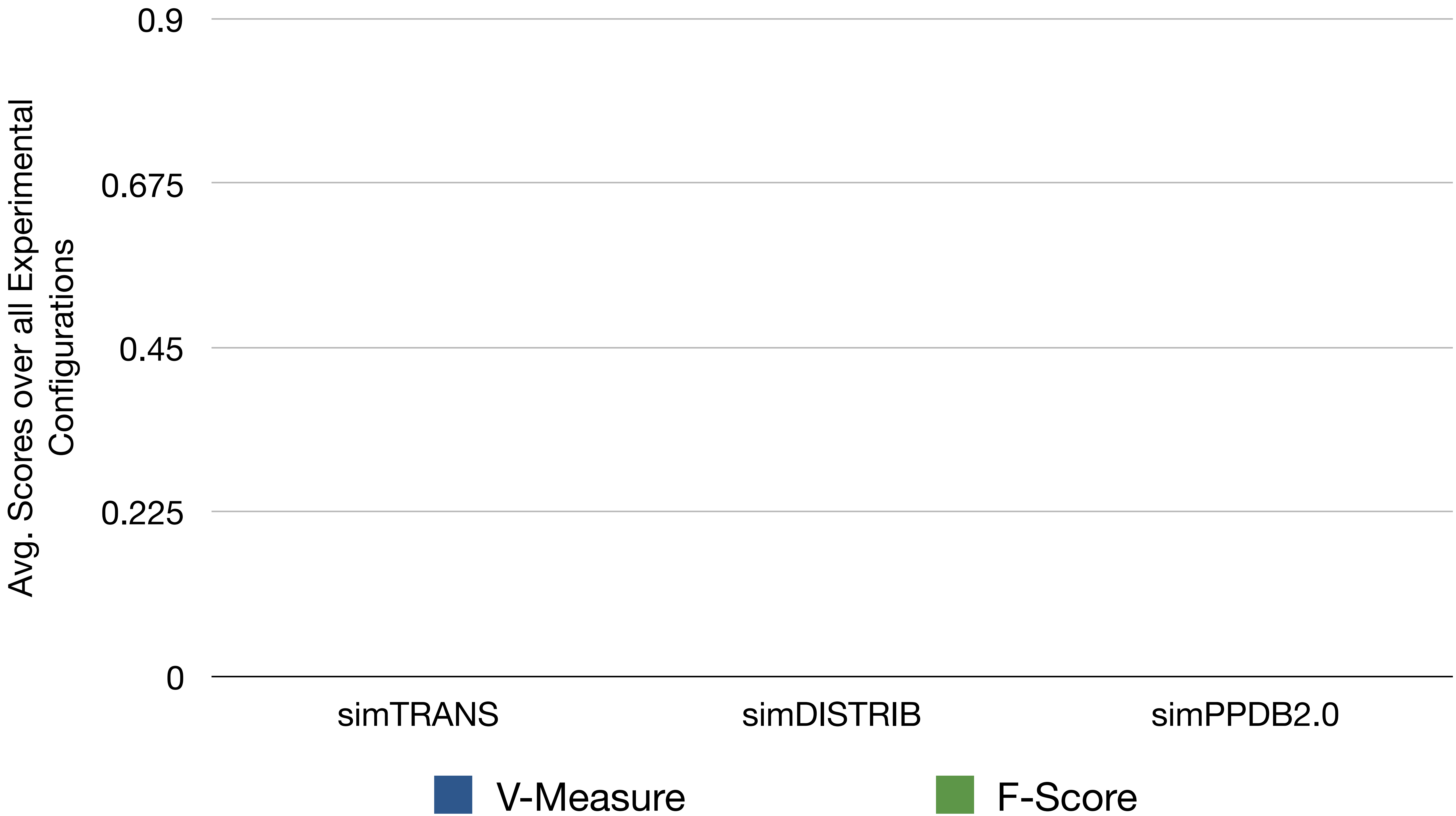


- CrowdClusters

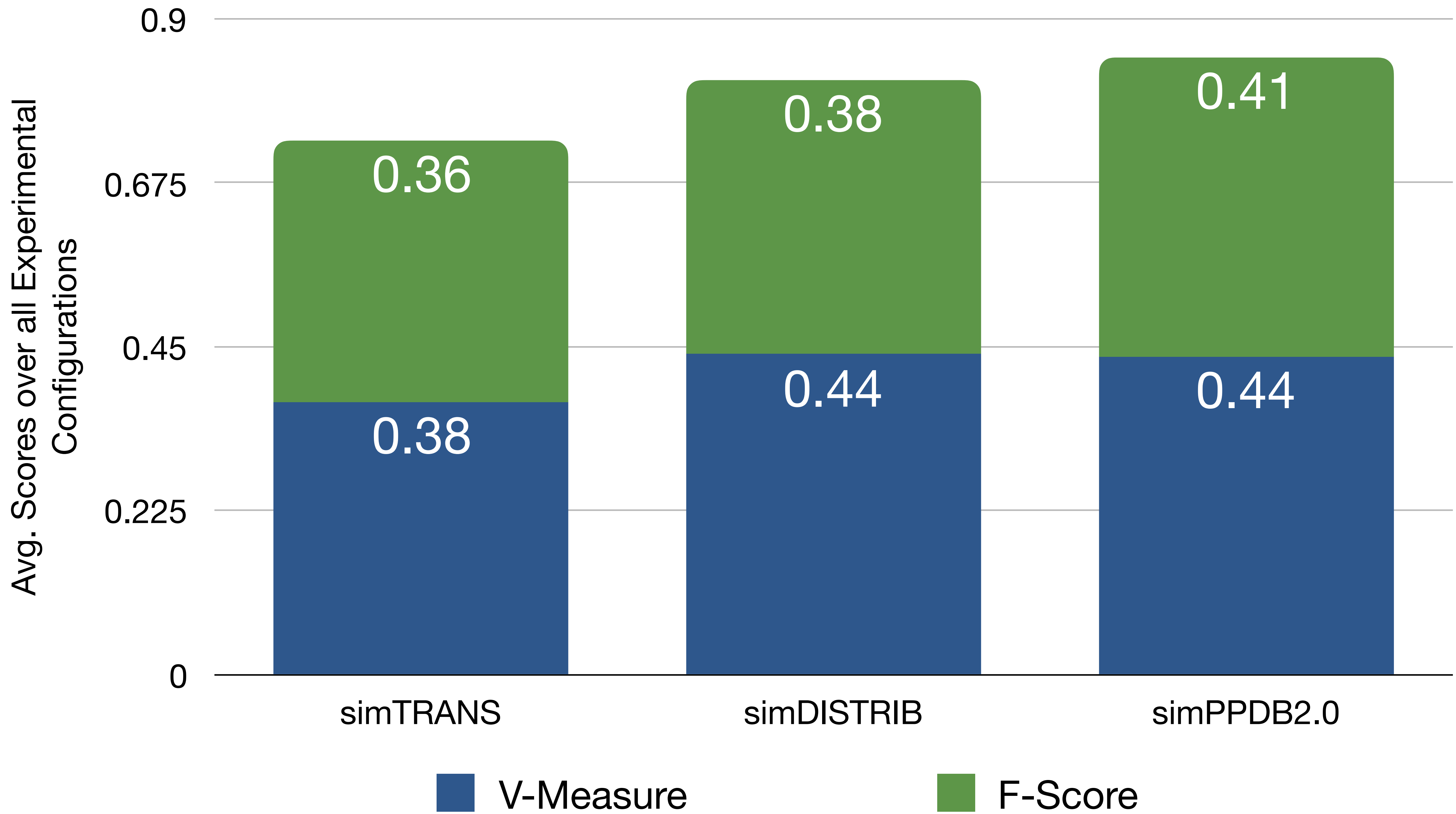


Clustering based on paraphrase strength out-performs other similarity measures on average

Clustering based on paraphrase strength out-performs other similarity measures on average

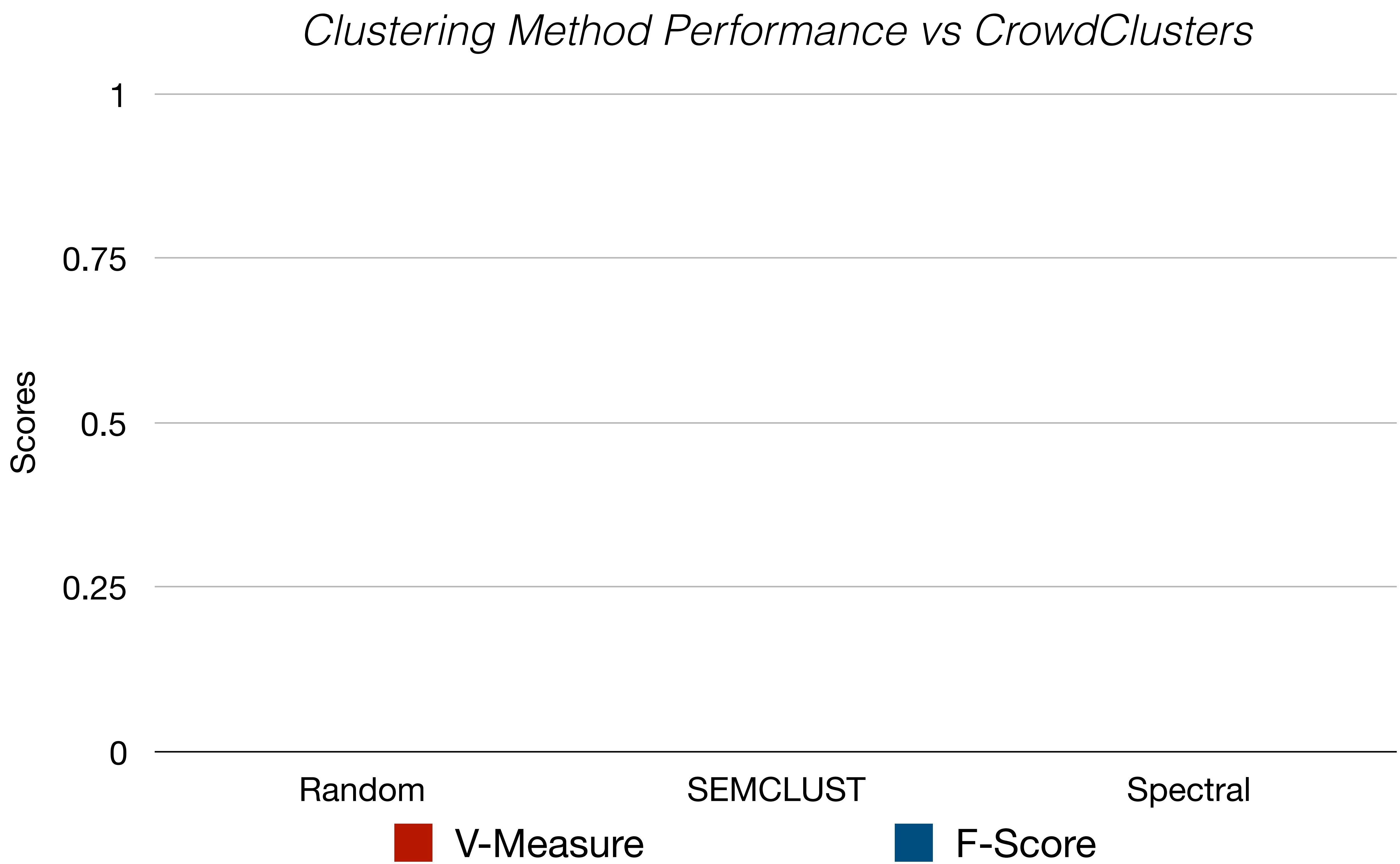


Clustering based on paraphrase strength out-performs other similarity measures on average

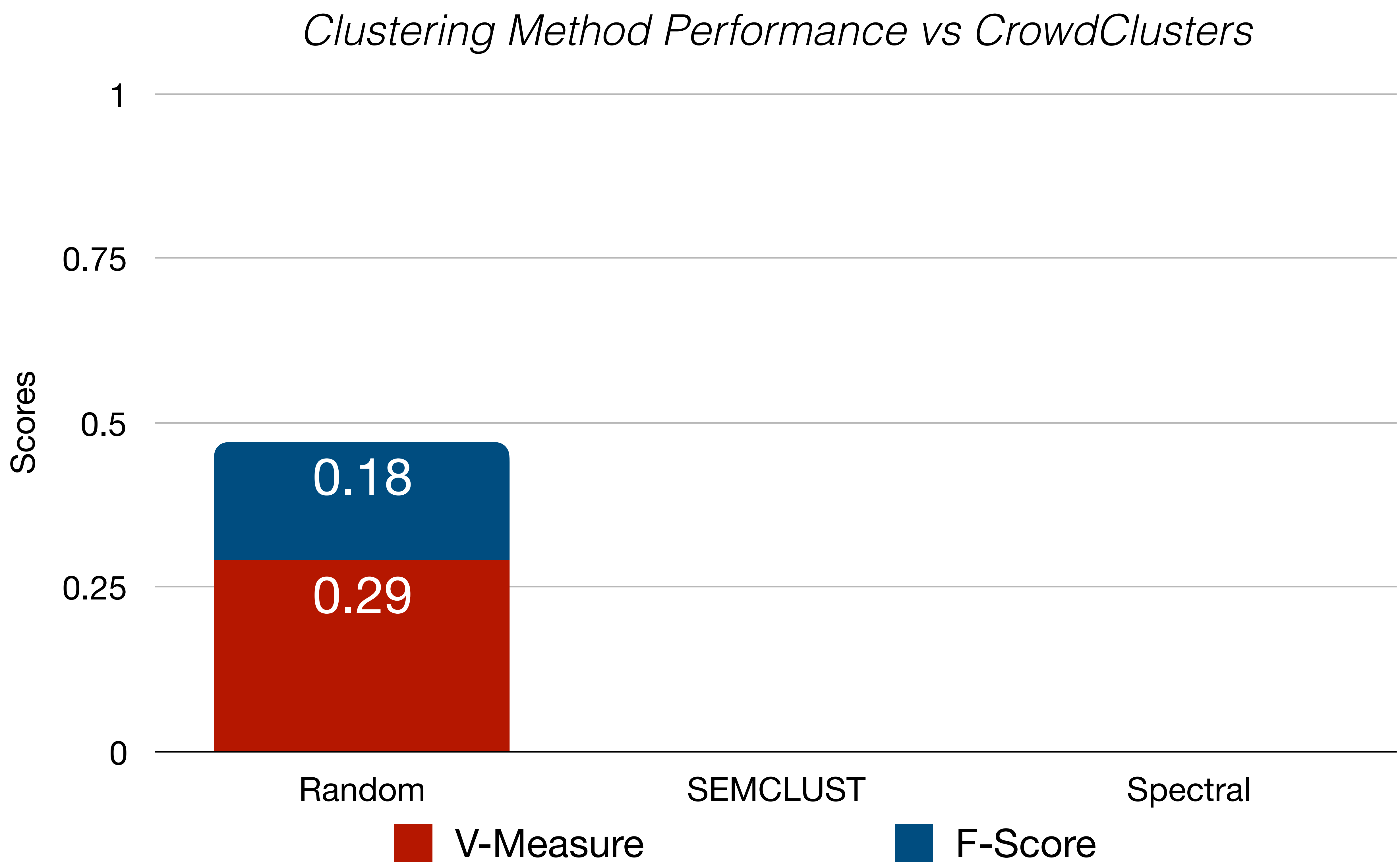


The best-performing spectral method uses $sim_{PPDB2.0}$ to compute similarity, and $sim_{DISTRIB}$ to estimate # clusters.

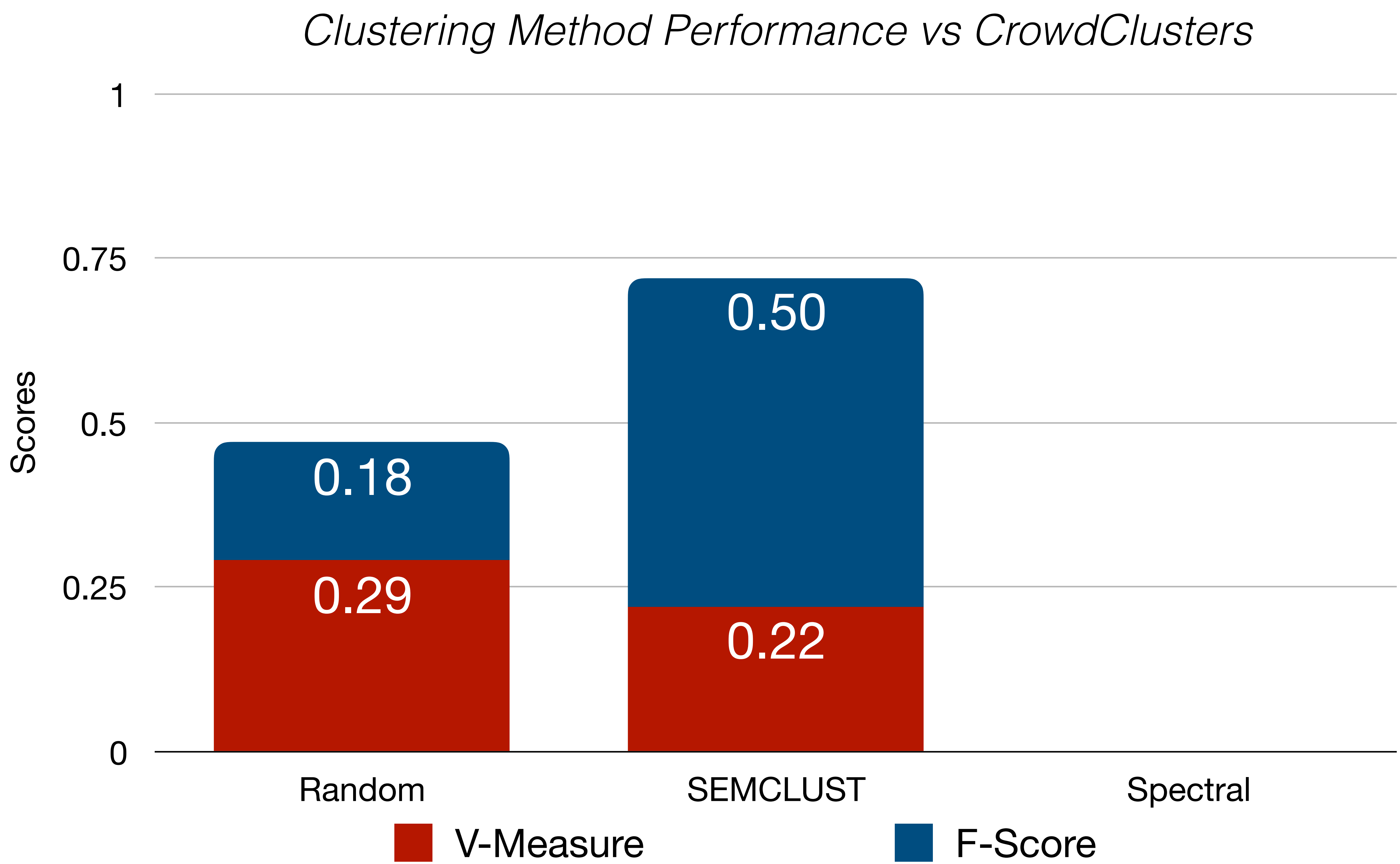
The best-performing spectral method uses $sim_{PPDB2.0}$ to compute similarity, and $sim_{DISTRIB}$ to estimate # clusters.



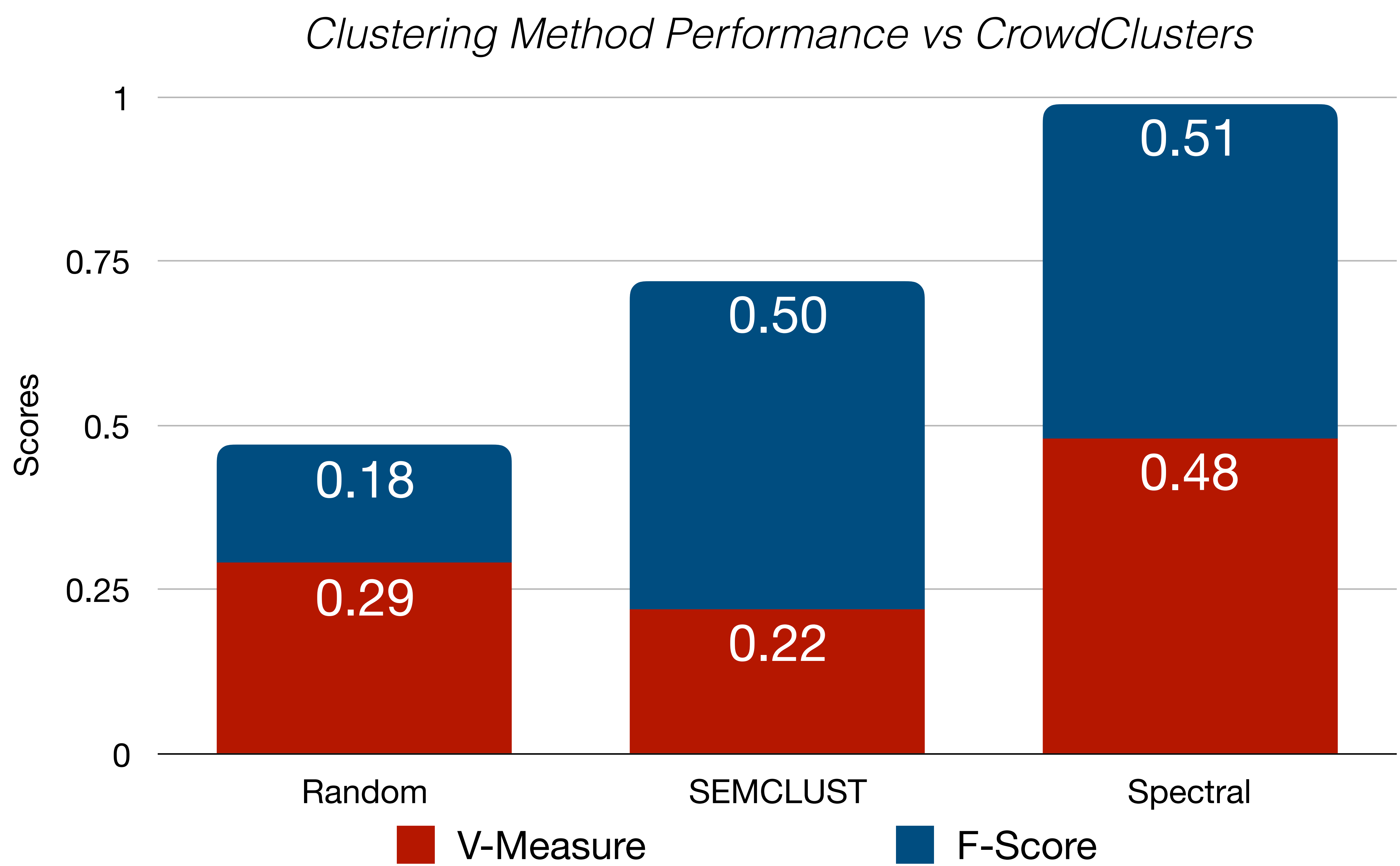
The best-performing spectral method uses $sim_{PPDB2.0}$ to compute similarity, and $sim_{DISTRIB}$ to estimate # clusters.



The best-performing spectral method uses $\text{sim}_{\text{PPDB2.0}}$ to compute similarity, and $\text{sim}_{\text{DISTRIB}}$ to estimate # clusters.

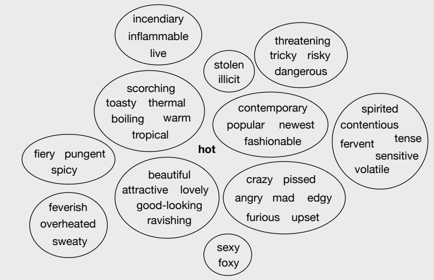


The best-performing spectral method uses $sim_{PPDB2.0}$ to compute similarity, and $sim_{DISTRIB}$ to estimate # clusters.



Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017



- Claims:
  - Paraphrases can be used to model the different meanings of a target word through *sense clustering*
  - The resulting *sense clusters* can be used to help find the most applicable substitutes for a target word in context

The **lexical substitution** task asks systems to propose appropriate substitutes for a target word in context

"There is a **bug** that causes the settings to get cleared."

The **lexical substitution** task asks systems to propose appropriate substitutes for a target word in context

"There is a bug that causes the settings to get cleared."

problem
error
issue
glitch

human-generated

The **lexical substitution** task asks systems to propose appropriate substitutes for a target word in context

"There is a bug that causes the settings to get cleared."

problem
error
issue
glitch

human-generated

virus
glitch
cookie
issue
error
problem
fault
insect
pest
rat
mole
tracker

system-generated

The **lexical substitution** task asks systems to propose appropriate substitutes for a target word in context

"There is a bug that causes the settings to get cleared."

problem
error
issue
glitch

human-generated

virus
glitch
cookie
issue
error
problem
fault
insect
pest
rat
mole
tracker

system-generated

$$p@1 = \frac{0}{1} = 0.0$$

The **lexical substitution** task asks systems to propose appropriate substitutes for a target word in context

"There is a **bug** that causes the settings to get cleared."

problem
error
issue
glitch

human-generated

virus
glitch
cookie
issue
error
problem
fault
insect
pest
rat
mole
tracker

system-generated

$$p@1 = \frac{0}{1} = 0.0$$

$$p@5 = \frac{2}{5} = 0.4$$

State-of-the-art systems propose substitutes based on word embeddings that encode distributional similarity

"There is a **bug** that causes the settings to get cleared."

State-of-the-art systems propose substitutes based on word embeddings that encode distributional similarity

"There is a **bug** that causes the settings to get cleared."

A Simple Word Embedding Model for Lexical Substitution, Melamud et al. NAACL 2015

State-of-the-art systems propose substitutes based on word embeddings that encode distributional similarity

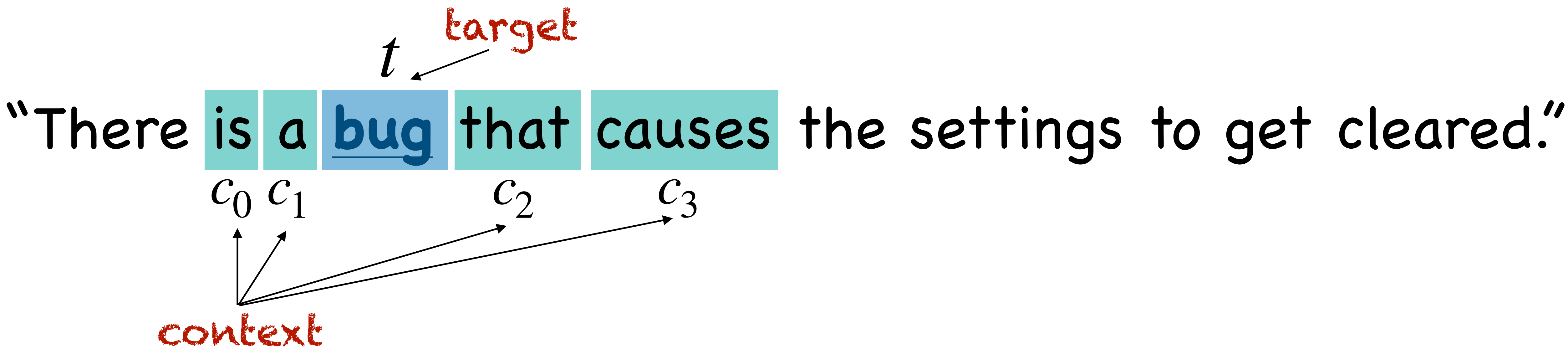
target

t

"There is a **bug** that causes the settings to get cleared."

A Simple Word Embedding Model for Lexical Substitution, Melamud et al. NAACL 2015

State-of-the-art systems propose substitutes based on word embeddings that encode distributional similarity



A Simple Word Embedding Model for Lexical Substitution, Melamud et al. NAACL 2015

State-of-the-art systems propose substitutes based on word embeddings that encode distributional similarity

target

t

"There is a bug that causes the settings to get cleared."

c_0 c_1 c_2 c_3

context

substitute → s `glitch`

A Simple Word Embedding Model for Lexical Substitution, Melamud et al. NAACL 2015

State-of-the-art systems propose substitutes based on word embeddings that encode distributional similarity

target $\rightarrow t$

"There is a bug that causes the settings to get cleared."

c_0 c_1 c_2 c_3

context

substitute $\rightarrow S$ **glitch**

$$AddCos(s, t, C) = \frac{cos(s,t) + \sum_{c \in C} cos(s,c)}{|C| + 1}$$

A Simple Word Embedding Model for Lexical Substitution, Melamud et al. NAACL 2015

State-of-the-art systems propose substitutes based on word embeddings that encode distributional similarity

target

t

"There is a bug that causes the settings to get cleared."

c_0 c_1 c_2 c_3

context

substitute $\rightarrow s$ **glitch**

$$AddCos(s, t, C) = \frac{cos(s, t) + \sum_{c \in C} cos(s, c)}{|C| + 1}$$

A Simple Word Embedding Model for Lexical Substitution, Melamud et al. NAACL 2015

State-of-the-art systems propose substitutes based on word embeddings that encode distributional similarity

target $\rightarrow t$

"There is a bug that causes the settings to get cleared."

$c_0 \quad c_1 \quad c_2 \quad c_3$

context

substitute $\rightarrow s$ glitch

$$AddCos(s, t, C) = \frac{cos(s, t) + \sum_{c \in C} cos(s, c)}{|C| + 1}$$

A Simple Word Embedding Model for Lexical Substitution, Melamud et al. NAACL 2015

Sense promotion elevates the rank of system-generated substitutes from the ‘best fit’ cluster

“There is a **bug** that causes the settings to get cleared.”

Sense promotion elevates the rank of system-generated substitutes from the 'best fit' cluster

"There is a **bug** that causes the settings to get cleared."

virus
glitch
error
wire
issue
problem
fault
insect
pest
rat
mole
tracker

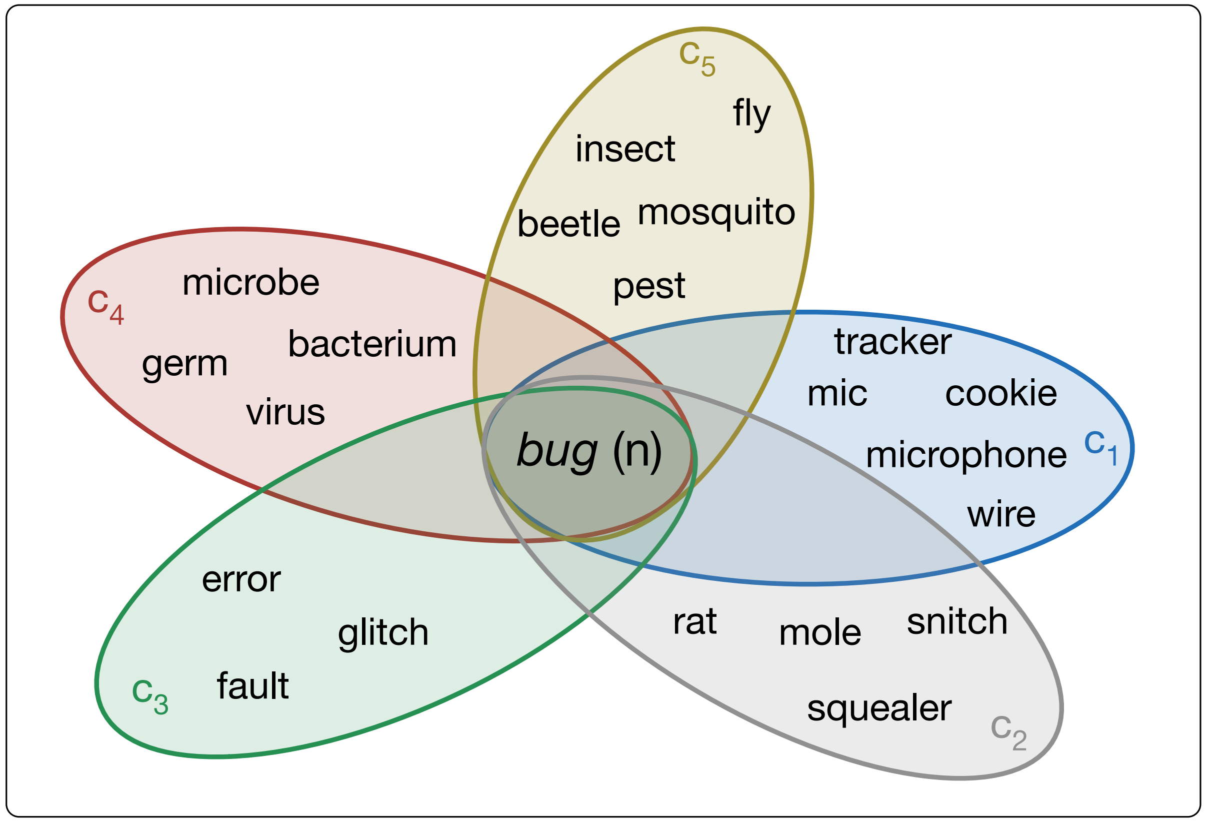
lexsub-system-generated

Sense promotion elevates the rank of system-generated substitutes from the 'best fit' cluster

"There is a **bug** that causes the settings to get cleared."

virus
glitch
error
wire
issue
problem
fault
insect
pest
rat
mole
tracker

lexsub-system-generated

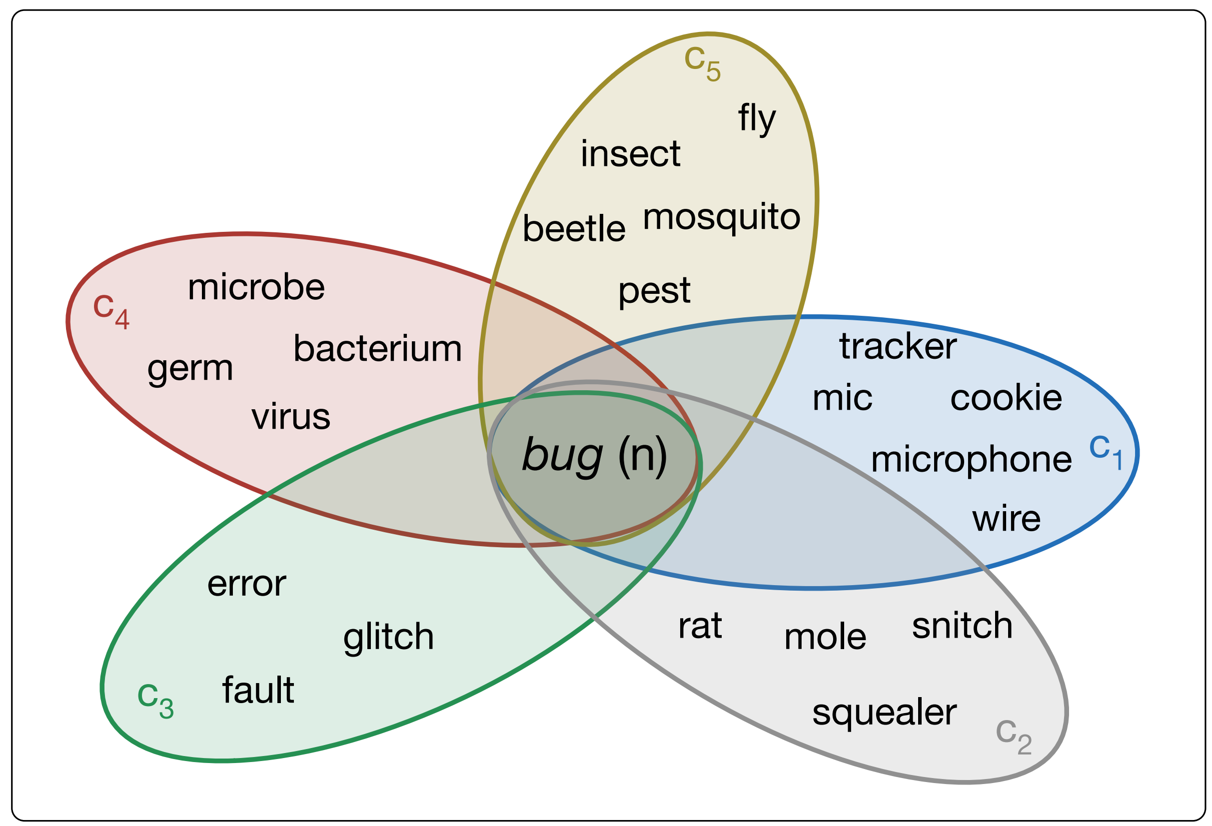


Sense promotion elevates the rank of system-generated substitutes from the 'best fit' cluster

"There is a **bug** that causes the settings to get cleared."

```
virus
glitch
error
wire
issue
problem
fault
insect
pest
rat
mole
tracker
```

lexsub-system-generated



error
tracker
fault
virus
glitch
insect
germ
cookie
mic
mole
squealer
...

Paraphrase Ranking for Choosing Cluster

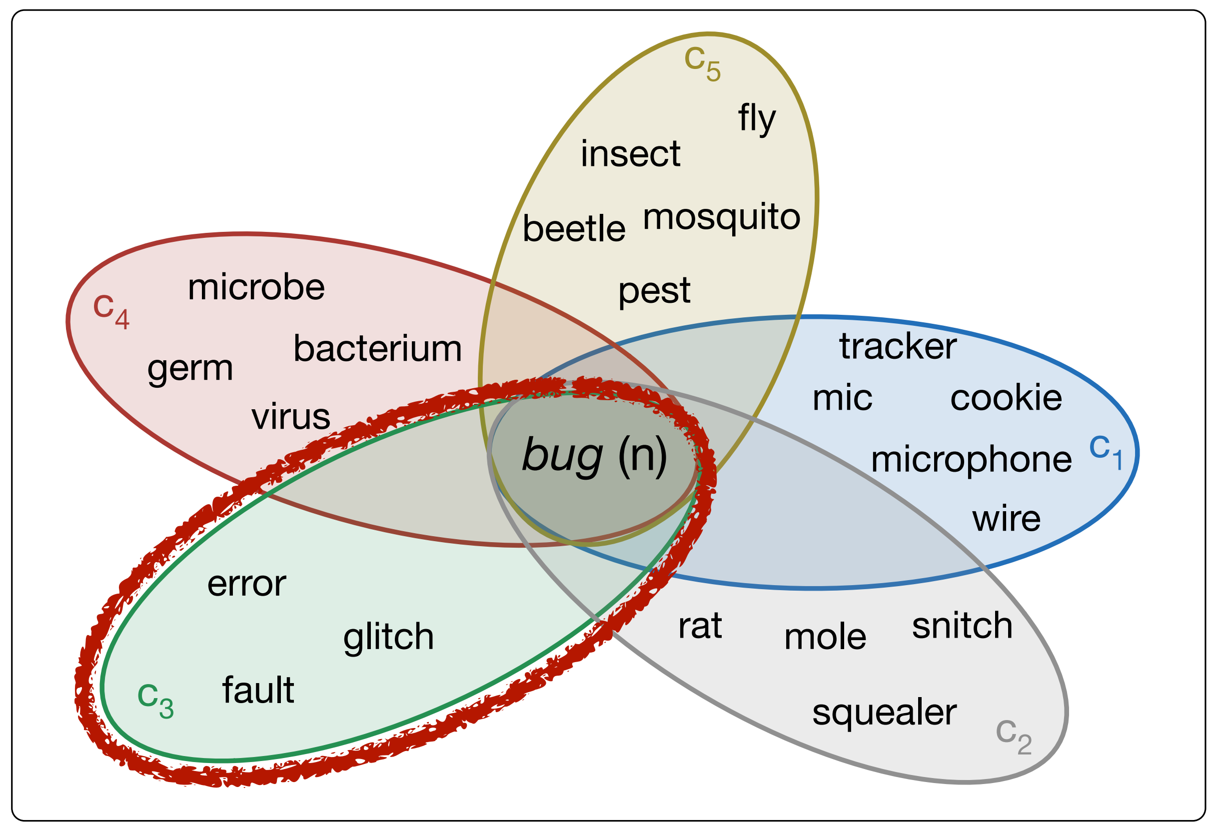
$$\sigma(p) \propto AddCos(p, t, C) \cdot PPDBScore(p, t)$$

Sense promotion elevates the rank of system-generated substitutes from the 'best fit' cluster

"There is a bug that causes the settings to get cleared."

virus
glitch
error
wire
issue
problem
fault
insect
pest
rat
mole
tracker

lexsub-system-generated



error
tracker
fault
virus
glitch
insect
germ
cookie
mic
mole
squealer
…

Paraphrase Ranking for Choosing Cluster

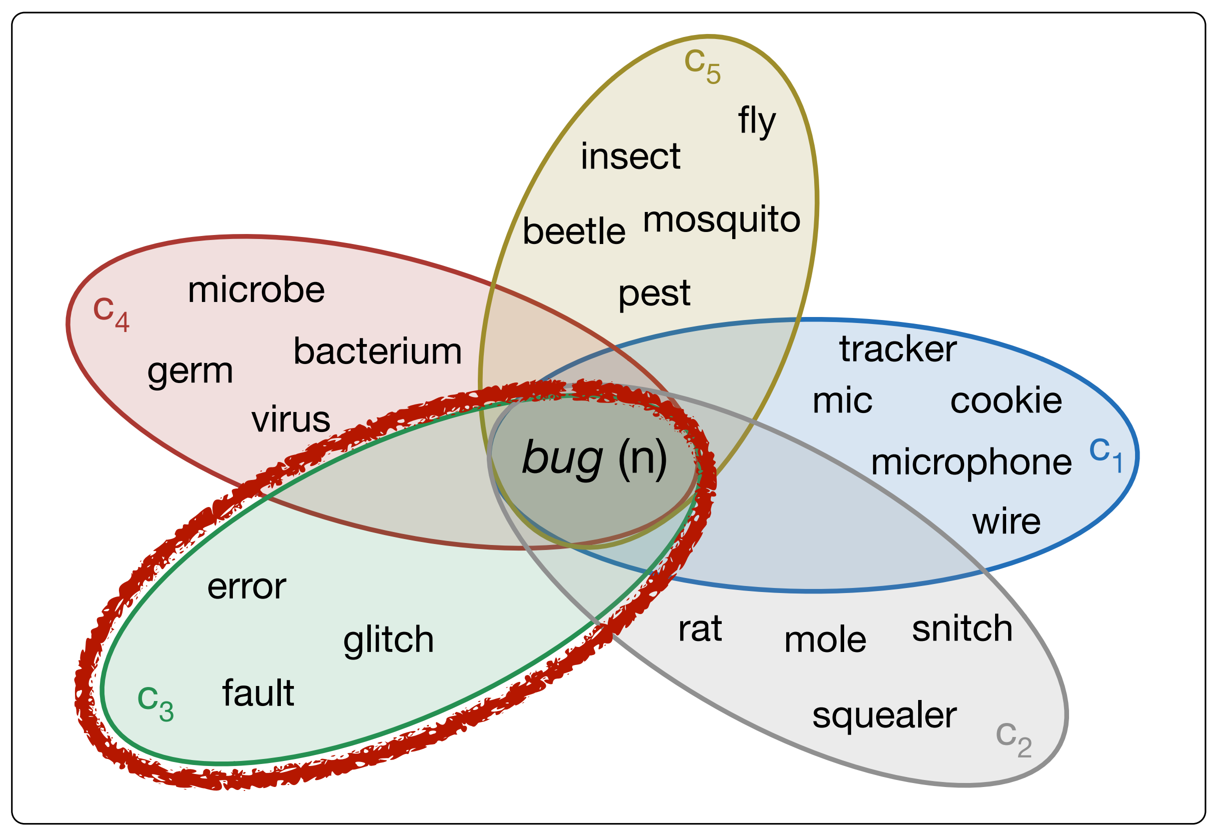
$$\sigma(p) \propto AddCos(p, t, C) \cdot PPDBScore(p, t)$$

Sense promotion elevates the rank of system-generated substitutes from the ‘best fit’ cluster

“There is a bug that causes the settings to get cleared.”

virus
glitch
error
wire
issue
problem
fault
insect
pest
rat
mole
tracker

lexsub-system-generated



error
tracker
fault
virus
glitch
insect
germ
cookie
mic
mole
squealer
…

Paraphrase Ranking for Choosing Cluster

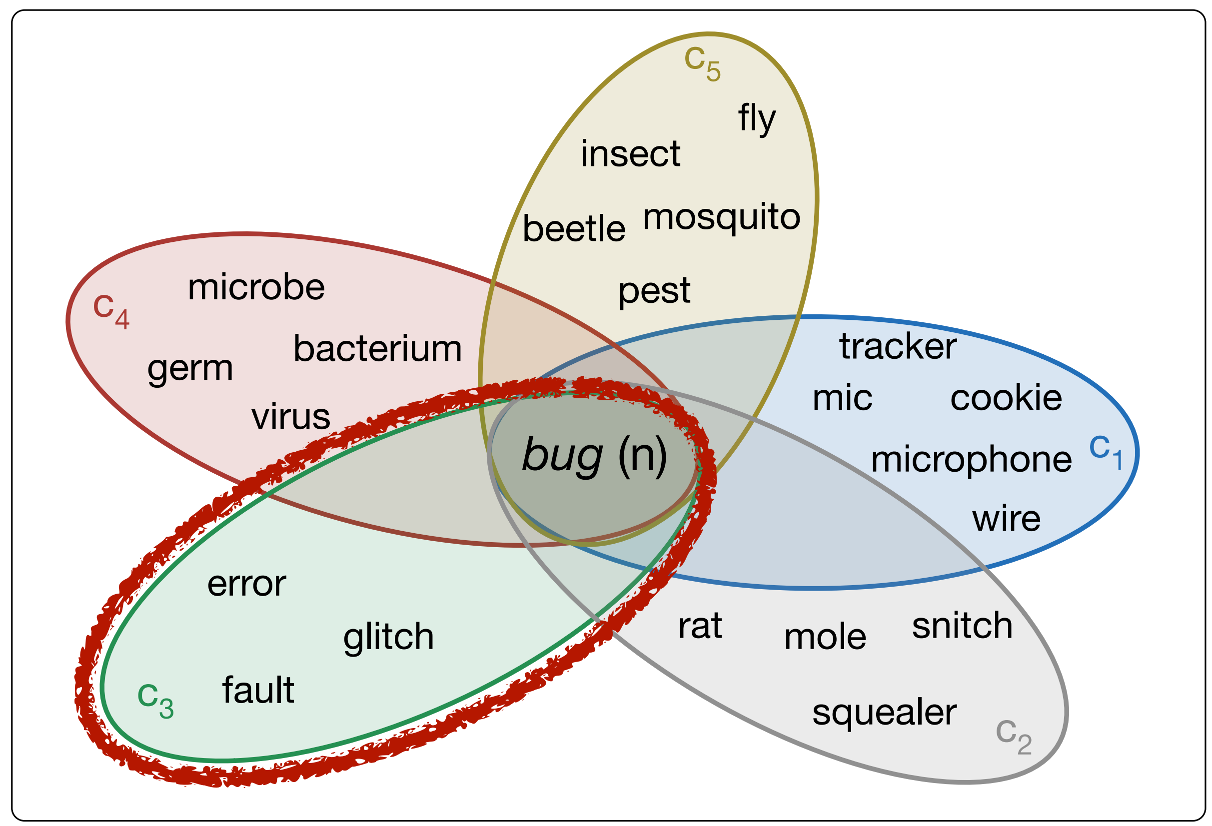
$$\sigma(p) \propto AddCos(p, t, C) \cdot PPDBScore(p, t)$$

Sense promotion elevates the rank of system-generated substitutes from the 'best fit' cluster

"There is a **bug** that causes the settings to get cleared."

glitch
error
fault
virus
wire
issue
problem
insect
pest
rat
mole
tracker

lexsub-system-generated



error
tracker
fault
virus
glitch
insect
germ
cookie
mic
mole
squealer
...

Paraphrase Ranking for Choosing Cluster

$$\sigma(p) \propto AddCos(p, t, C) \cdot PPDBScore(p, t)$$

Our experiments compare lexical substitution performance for two lexsub models, before and after sense promotion

Our experiments compare lexical substitution performance for two lexsub models, before and after sense promotion

- Dataset:

Our experiments compare lexical substitution performance for two lexsub models, before and after sense promotion

- Dataset:
    - Concepts in Context (CoInCo) — 2241 instances

Our experiments compare lexical substitution performance for two lexsub models, before and after sense promotion

- Dataset:
  - Concepts in Context (CoInCo) — 2241 instances
- For each target instance, we rank PPDB paraphrases of the target word as potential substitutes

Our experiments compare lexical substitution performance for two lexsub models, before and after sense promotion

- Dataset:
    - Concepts in Context (CoInCo) — 2241 instances
- For each target instance, we rank PPDB paraphrases of the target word as potential substitutes
- Lexsub Ranking Models:

Our experiments compare lexical substitution performance for two lexsub models, before and after sense promotion

- Dataset:
    - Concepts in Context (CoInCo) — 2241 instances
- For each target instance, we rank PPDB paraphrases of the target word as potential substitutes
- Lexsub Ranking Models:
    - AddCos (Melamud et al. 2015)

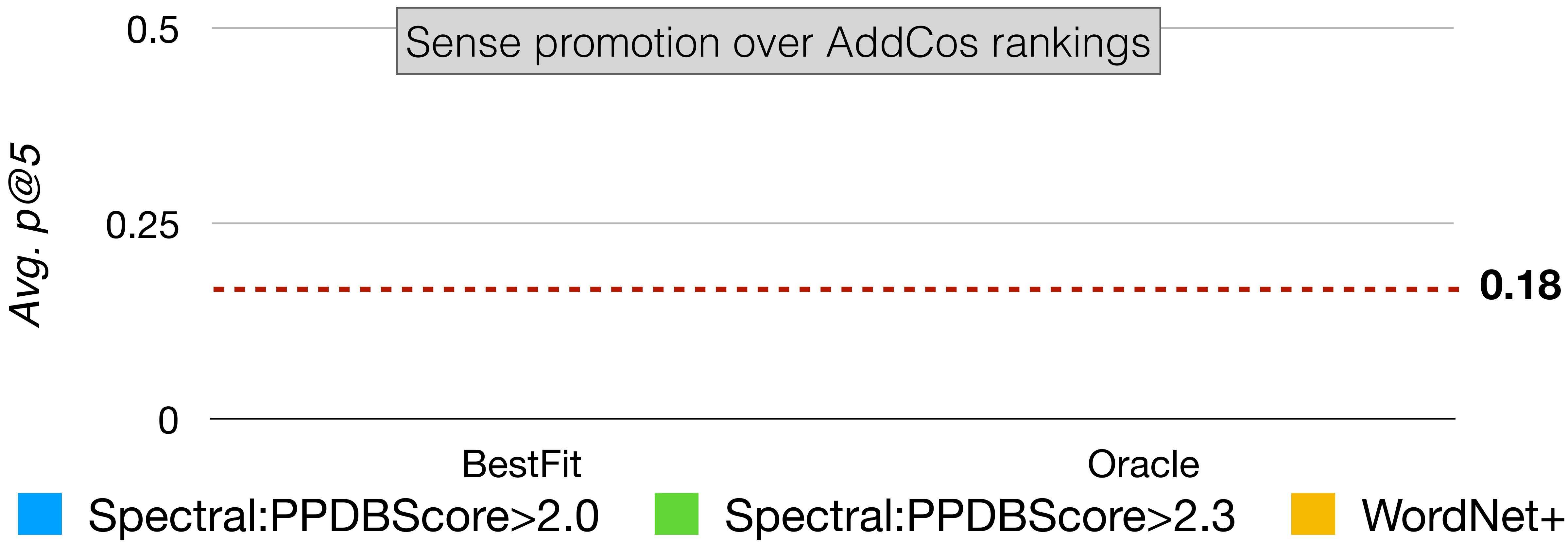
Our experiments compare lexical substitution performance for two lexsub models, before and after sense promotion

- Dataset:
  - Concepts in Context (CoInCo) — 2241 instances
- For each target instance, we rank PPDB paraphrases of the target word as potential substitutes
- Lexsub Ranking Models:
  - AddCos (Melamud et al. 2015)
  - Context2Vec (Melamud et al. 2016)

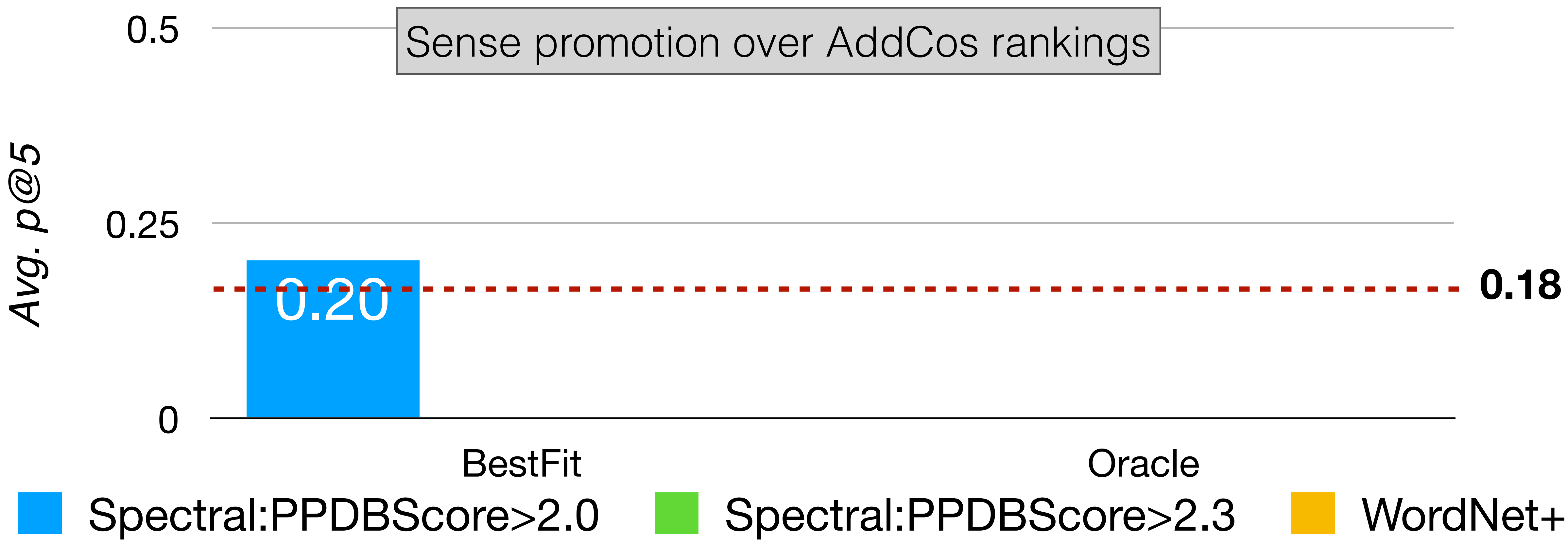
Sense promotion improves the precision of AddCos and Context2Vec rankings

Sense promotion over AddCos rankings

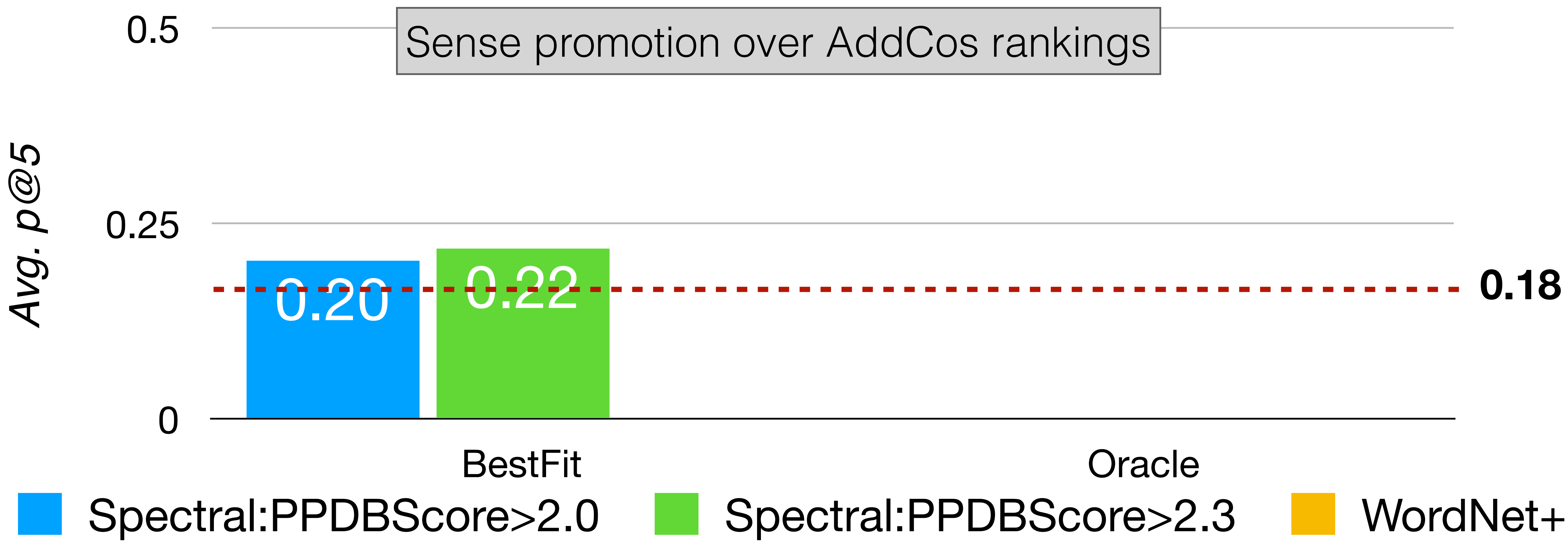
Sense promotion improves the precision of AddCos and Context2Vec rankings



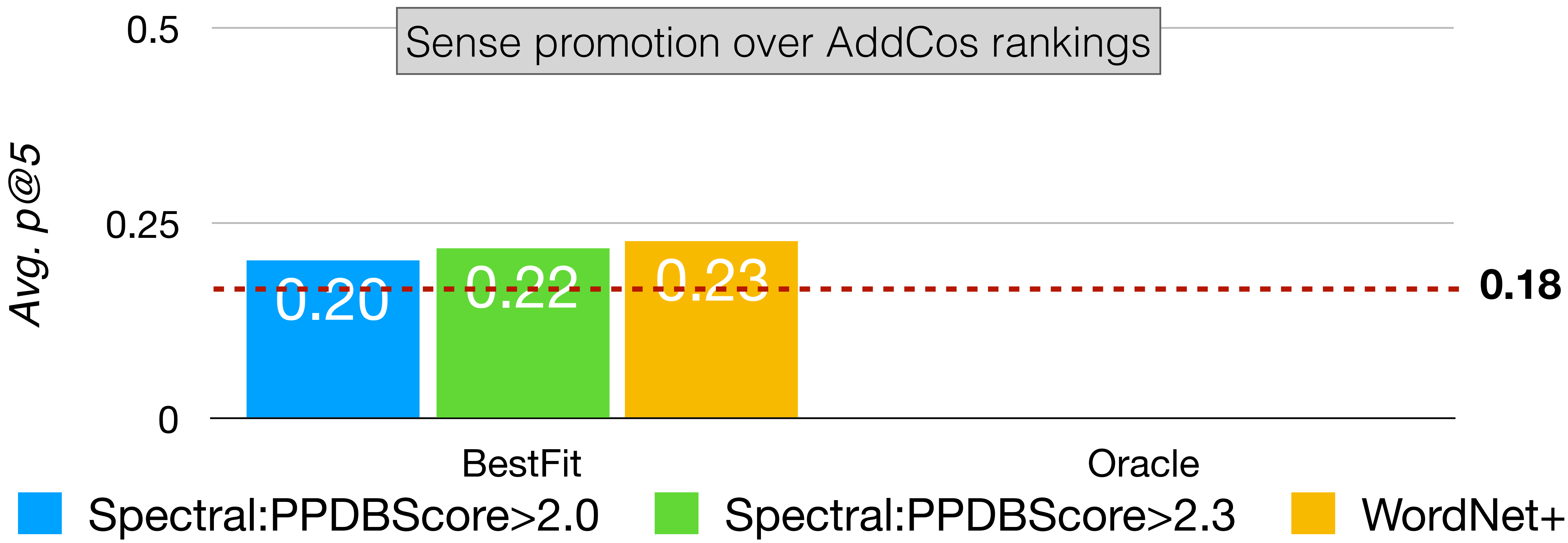
Sense promotion improves the precision of AddCos and Context2Vec rankings



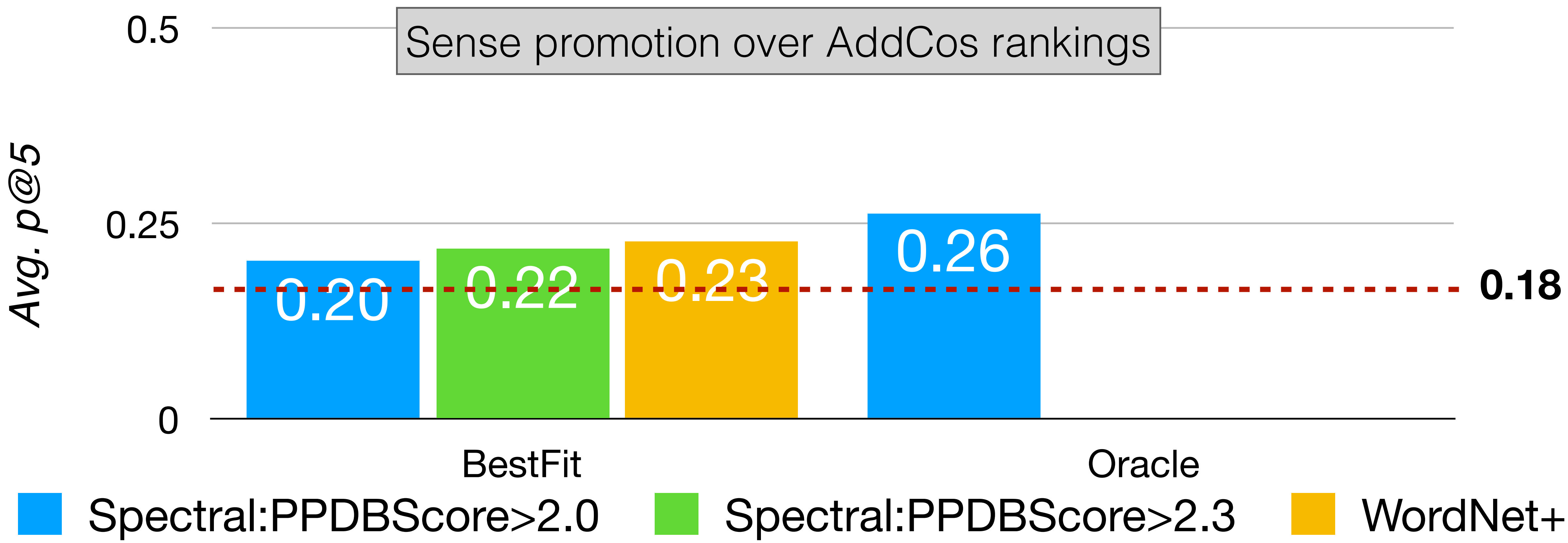
Sense promotion improves the precision of AddCos and Context2Vec rankings



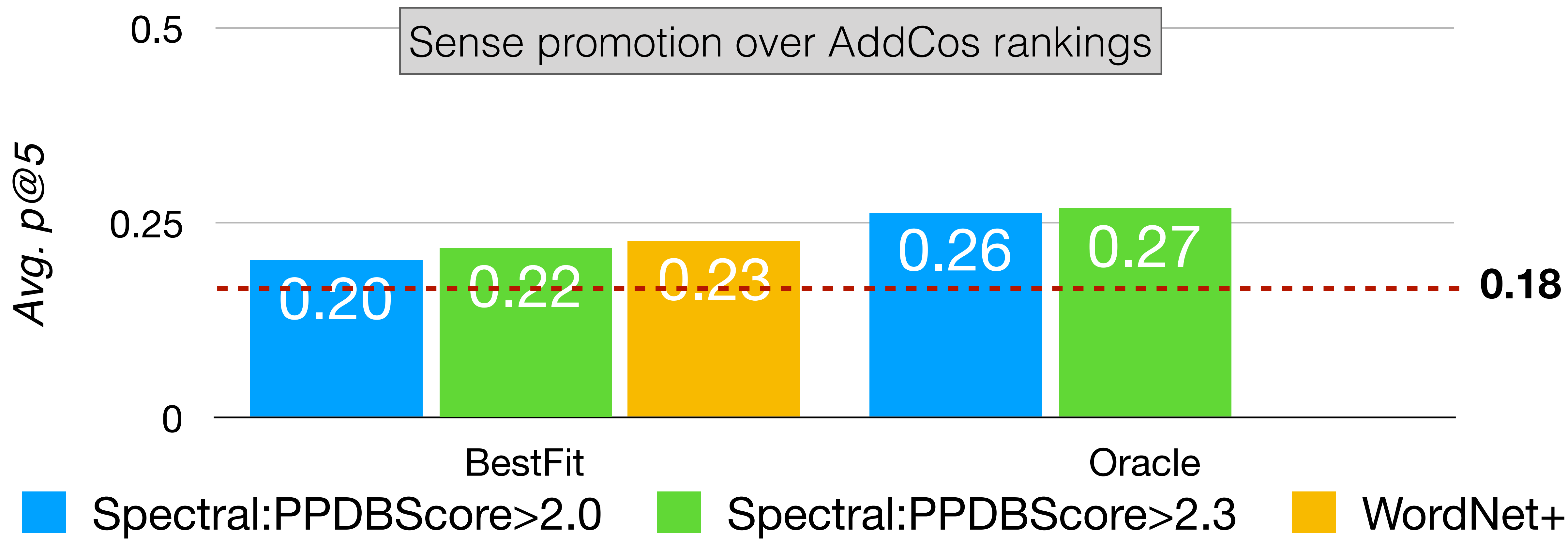
Sense promotion improves the precision of AddCos and Context2Vec rankings



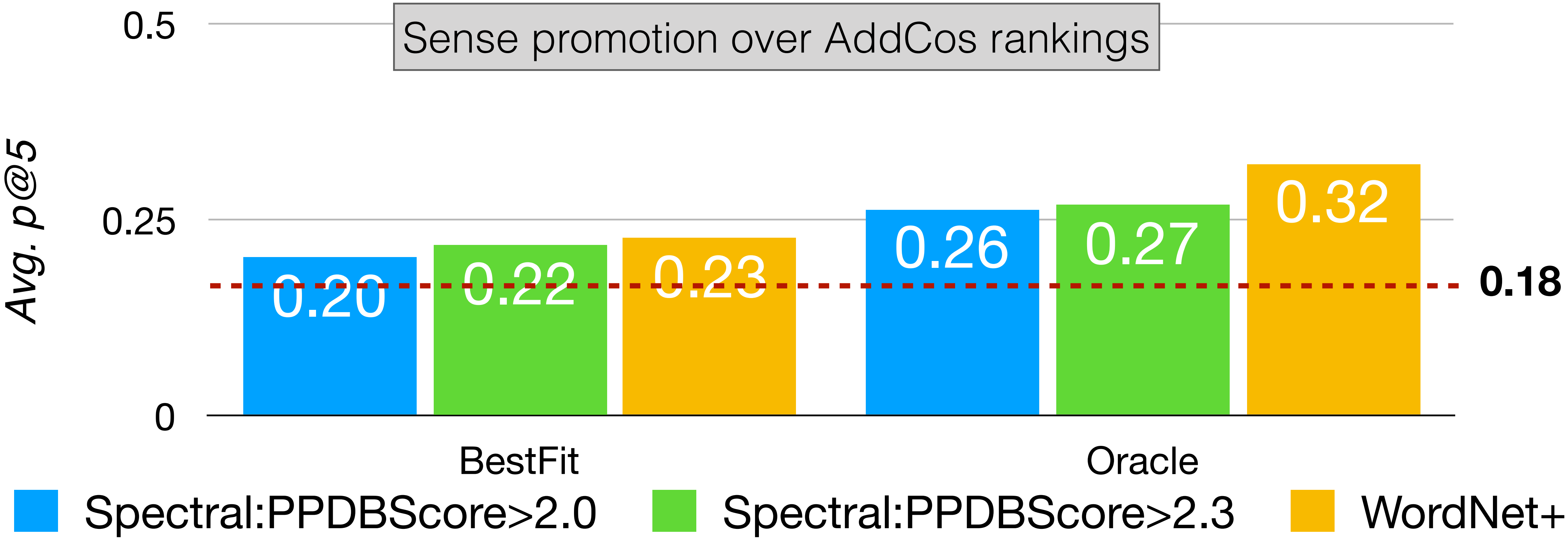
Sense promotion improves the precision of AddCos and Context2Vec rankings



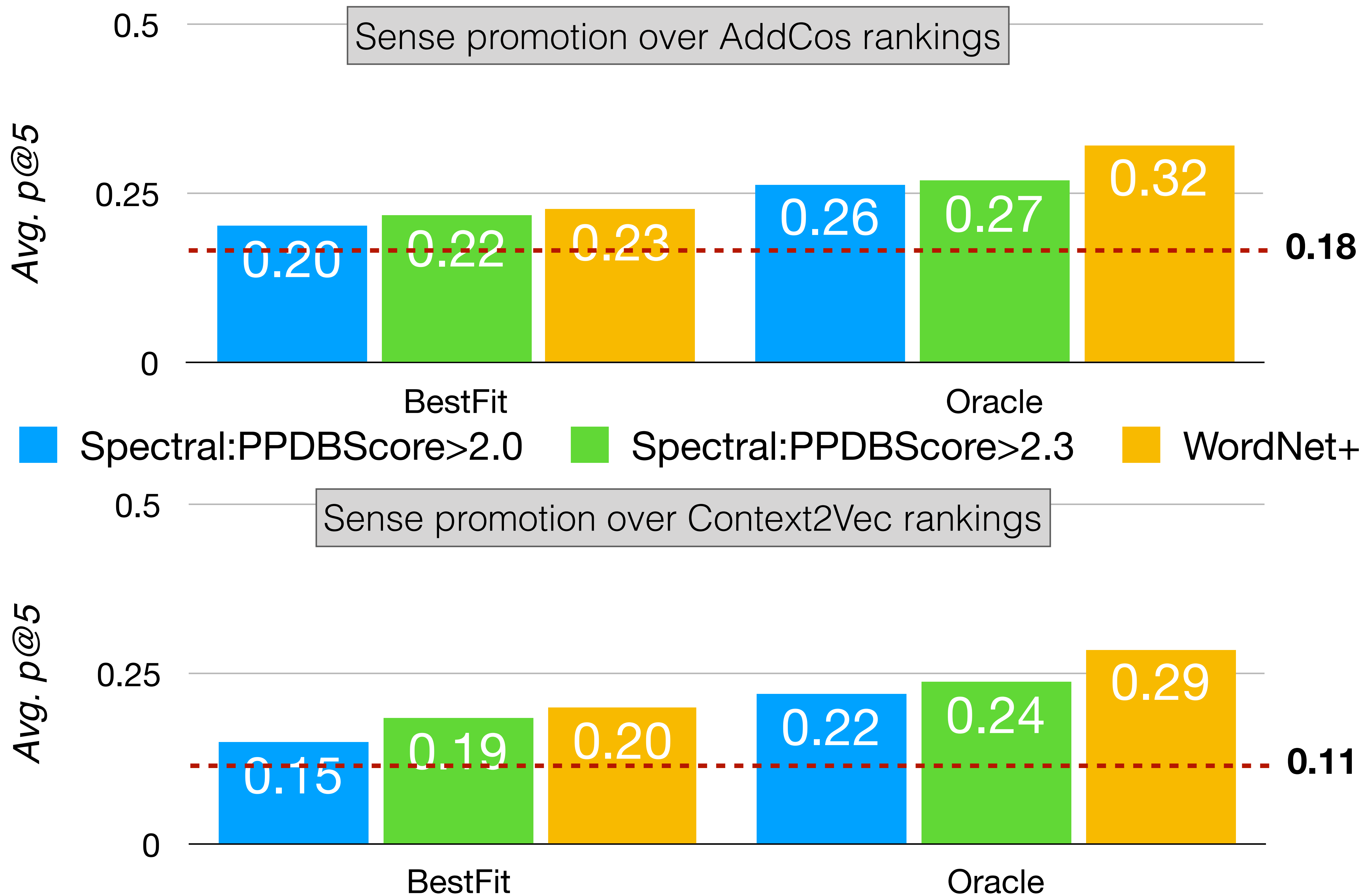
Sense promotion improves the precision of AddCos and Context2Vec rankings



Sense promotion improves the precision of AddCos and Context2Vec rankings



Sense promotion improves the precision of AddCos and Context2Vec rankings



Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

- Claims:
  - Paraphrases can be used to model the different meanings of a target word through *sense clustering*
  - The resulting *sense clusters* can be used to help find the most applicable substitutes for a target word in context

Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

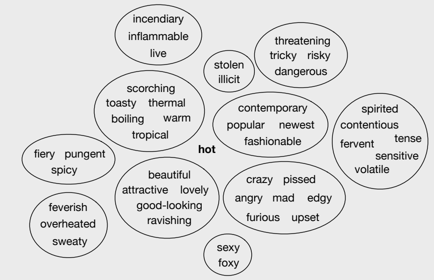
Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

- Take-aways:

Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017



- Take-aways:
    - Paraphrase strength is a useful signal for discriminating between different word meanings within a paraphrase set

Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

- Take-aways:
  - Paraphrase strength is a useful signal for discriminating between different word meanings within a paraphrase set
  - Best sense distinctions are made by combining paraphrase strength with distributional similarity signals

Motivation

Using Paraphrases to Model Word Sense

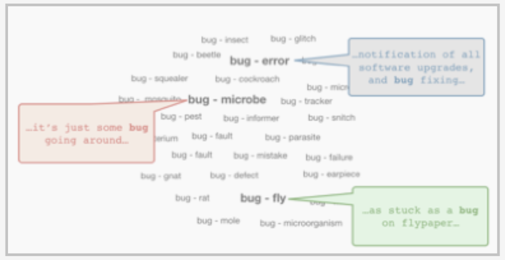
NAACL 2016; SENSE@EACL 2017

Learning Scalar Adjective Intensity

EMNLP 2018

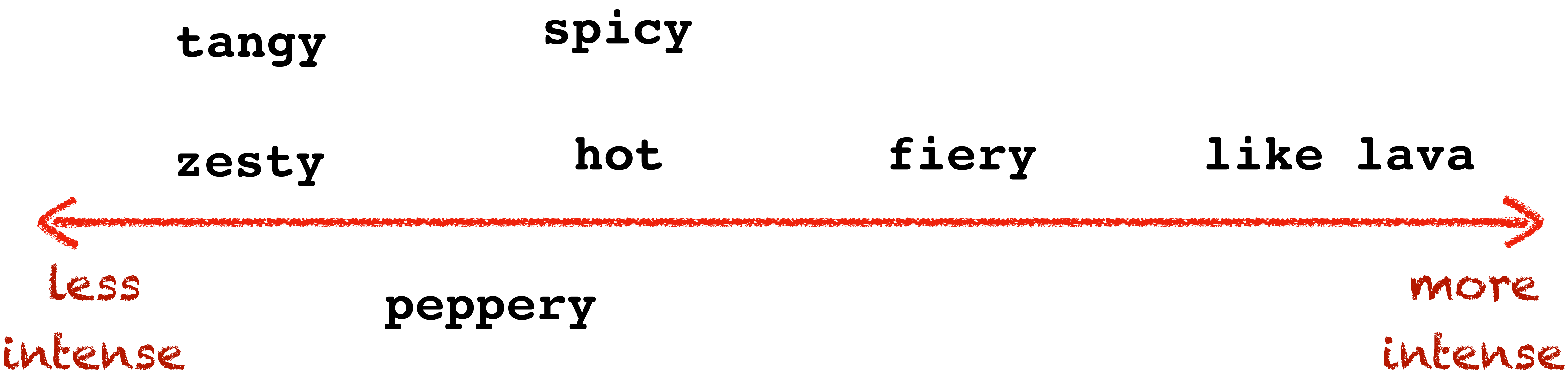
Meaning-specific Examples of Word Use

In submission



Conclusion

"What's a Chinese dish that's **not too hot**?"



Learning Scalar Adjective Intensity

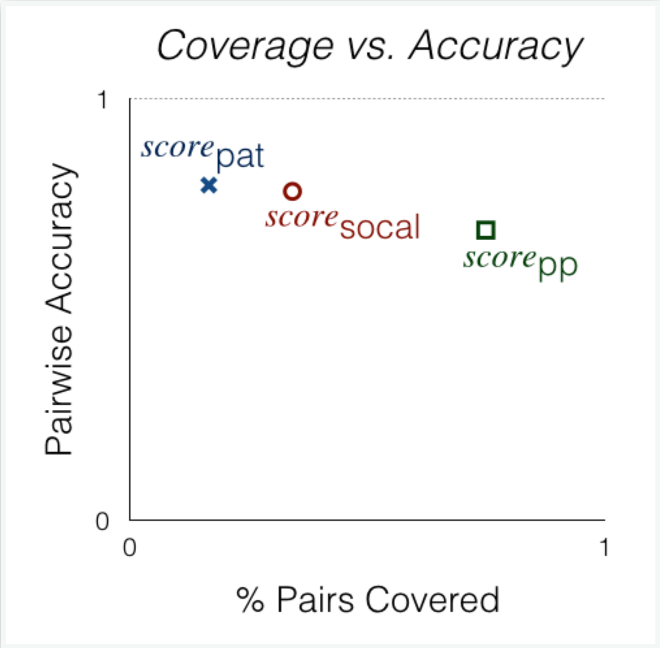
EMNLP 2018

hot < fiery

- Claims:
  - We can use adjectival phrase paraphrases to predict relative adjective intensity

really hot <--> fiery

  - This paraphrase-based information is complementary to pattern- and lexicon-based information



Adjectival paraphrases give evidence of relative adjective intensity

Paraphrase pair…			…is evidence that
particularly pleased	↔	*ecstatic*	*pleased* < *ecstatic*
quite limited	↔	*restricted*	*limited* < *restricted*
rather odd	↔	*crazy*	*odd* < *crazy*
so silly	↔	*dumb*	*silly* < *dumb*
completely mad	↔	*crazy*	*mad* < *crazy*

Adjectival paraphrases give evidence of relative adjective intensity

Paraphrase pair…			…is evidence that
particularly pleased	⟷	*ecstatic*	*pleased* < *ecstatic*
quite limited	⟷	*restricted*	*limited* < *restricted*
rather odd	⟷	*crazy*	*odd* < *crazy*
so silly	⟷	*dumb*	*silly* < *dumb*
completely mad	⟷	*crazy*	*mad* < *crazy*
RB JJ_1	⟷	JJ_2	$JJ_1 < JJ_2$

intensifying adverb

Using paraphrase-based signals to predict relative adjective intensity

- Challenge 1: Identify intensifying adverbs

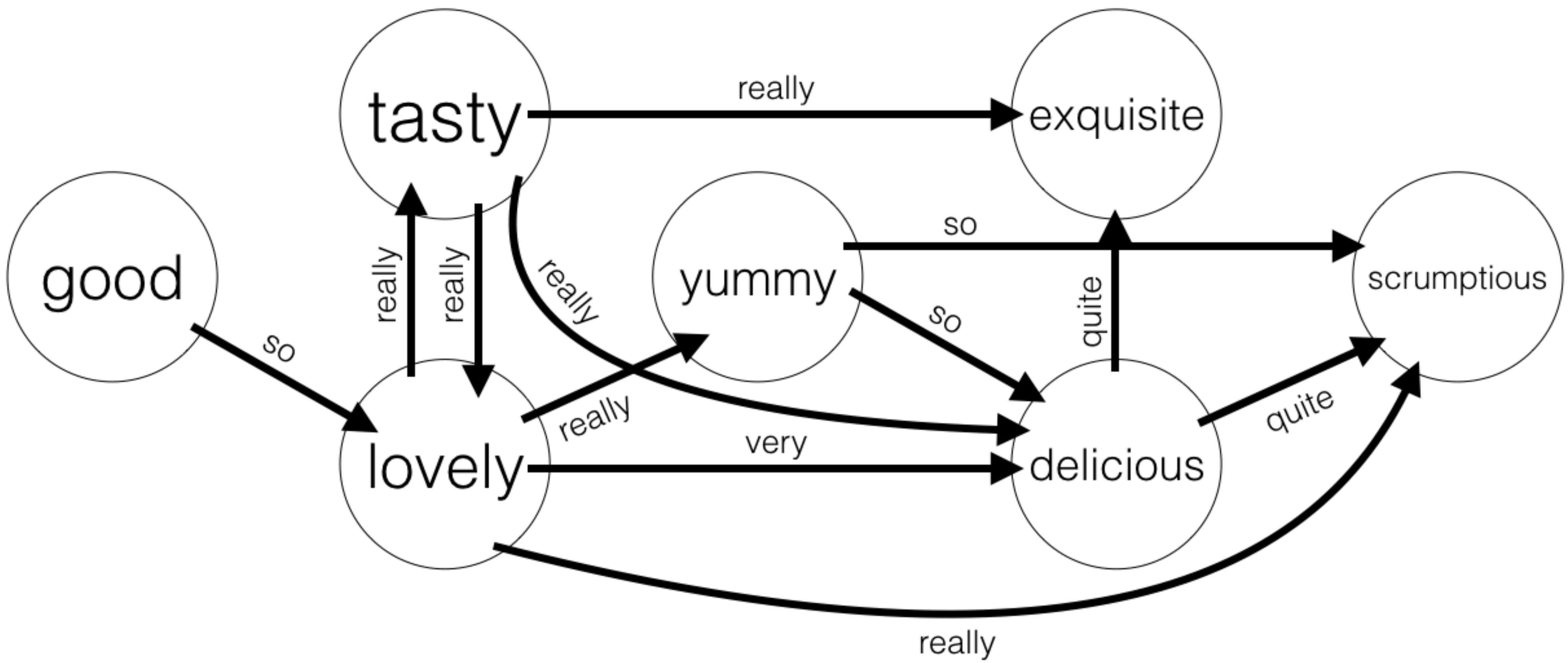
Using paraphrase-based signals to predict relative adjective intensity

- Challenge 1: Identify intensifying adverbs

hard < harder
harder < hardest
hard < hardest

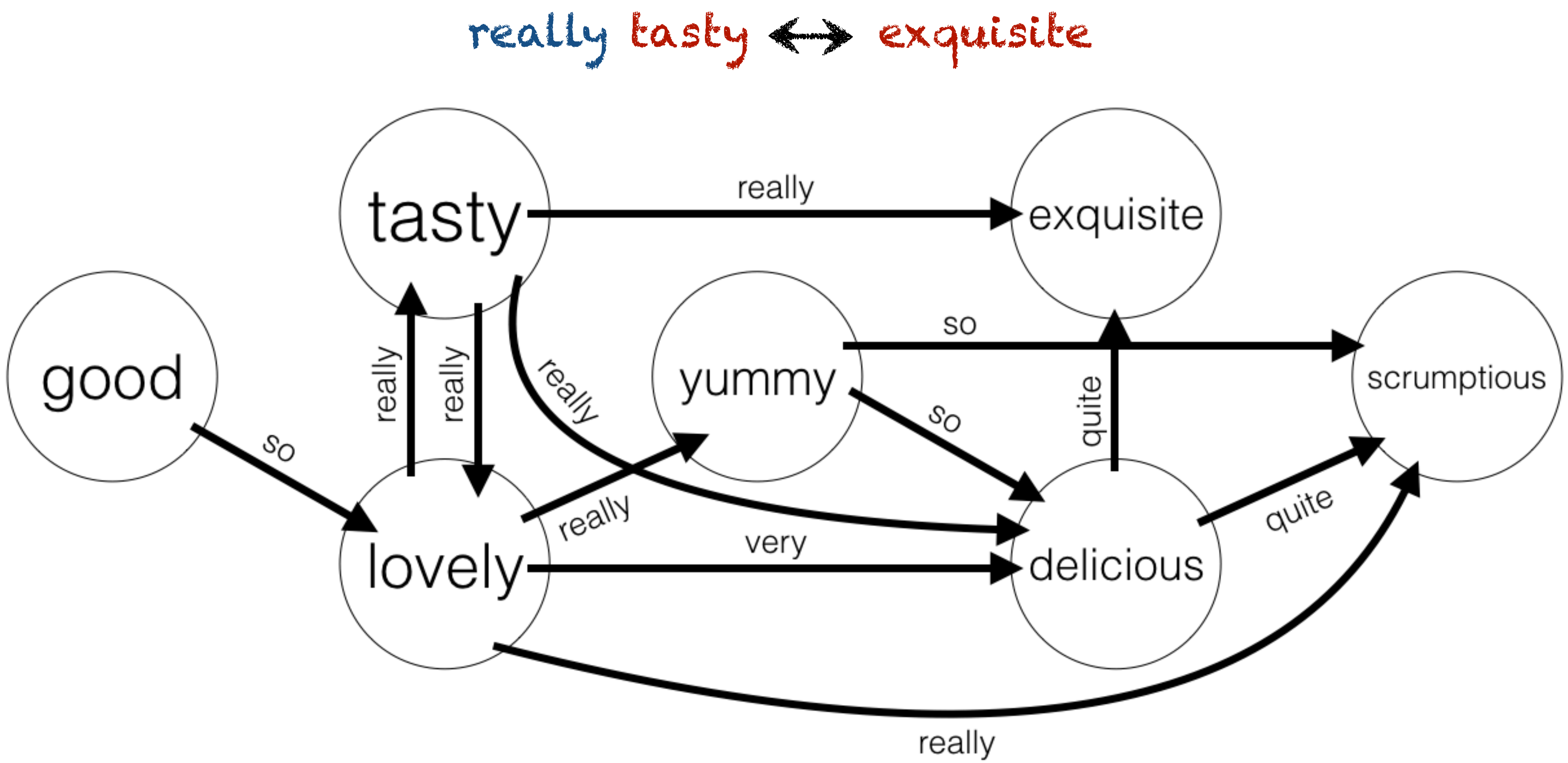
Round 1	**very**	hard	$\leftrightarrow$	harder
	kinda	hard	$\leftrightarrow$	harder
	so	hard	$\leftrightarrow$	harder
	pretty	hard	$\leftrightarrow$	harder
Round 2	very	*pleasant*	$\leftrightarrow$	*delightful*
	kinda	*hard*	$\leftrightarrow$	*tricky*
	so	*wonderful*	$\leftrightarrow$	*brilliant*
	pretty	*simple*	$\leftrightarrow$	*plain*
Round 3	**more**	pleasant	$\leftrightarrow$	delightful
	really	hard	$\leftrightarrow$	tricky
	truly	wonderful	$\leftrightarrow$	brilliant
	quite	simple	$\leftrightarrow$	plain

A graph structure can be used to encode the PPDB paraphrases matching our template



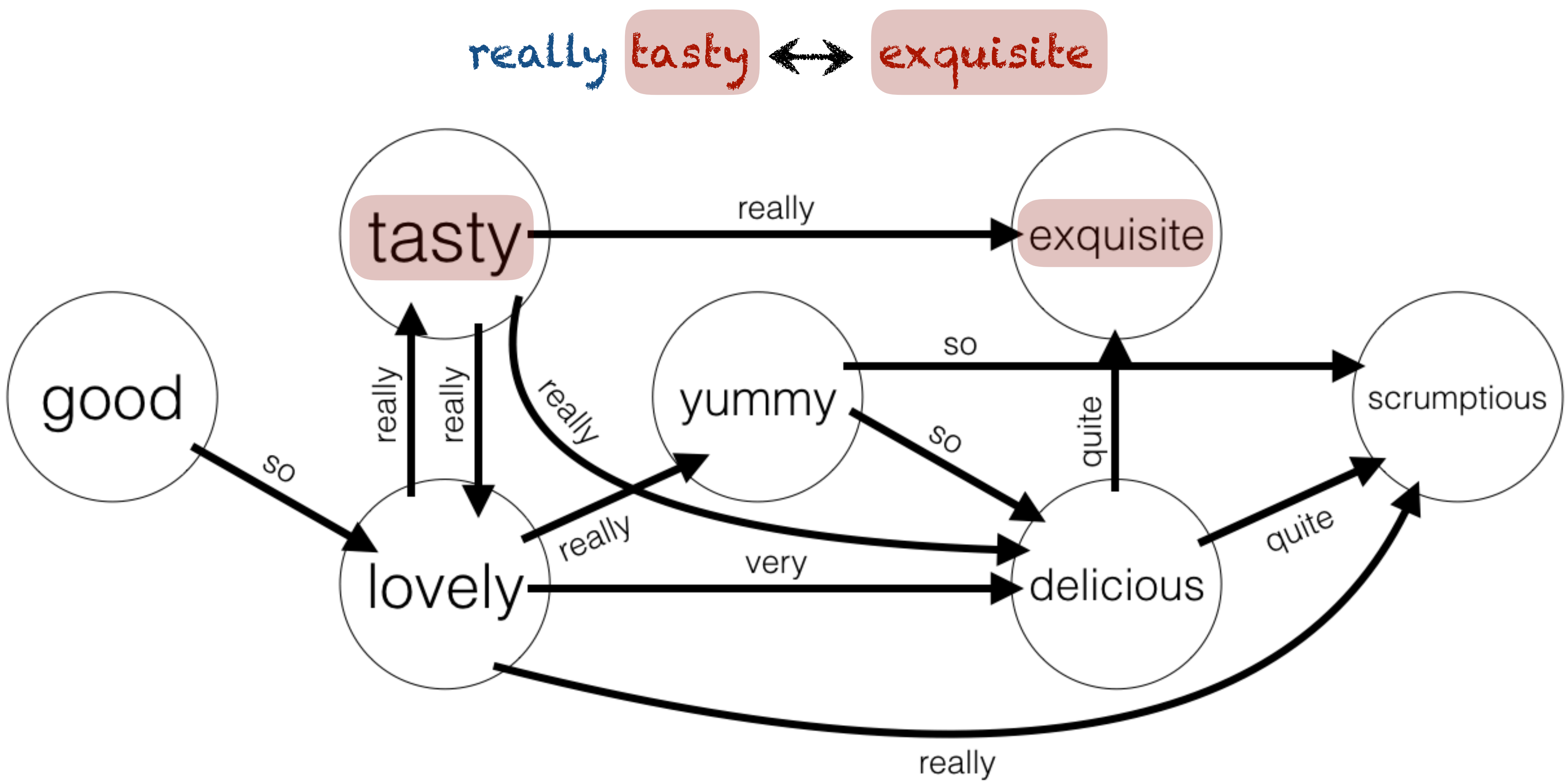
“JJGraph”: 3,704 nodes (adjectives) and ~36k edges (610 unique adverb labels)

A graph structure can be used to encode the PPDB paraphrases matching our template



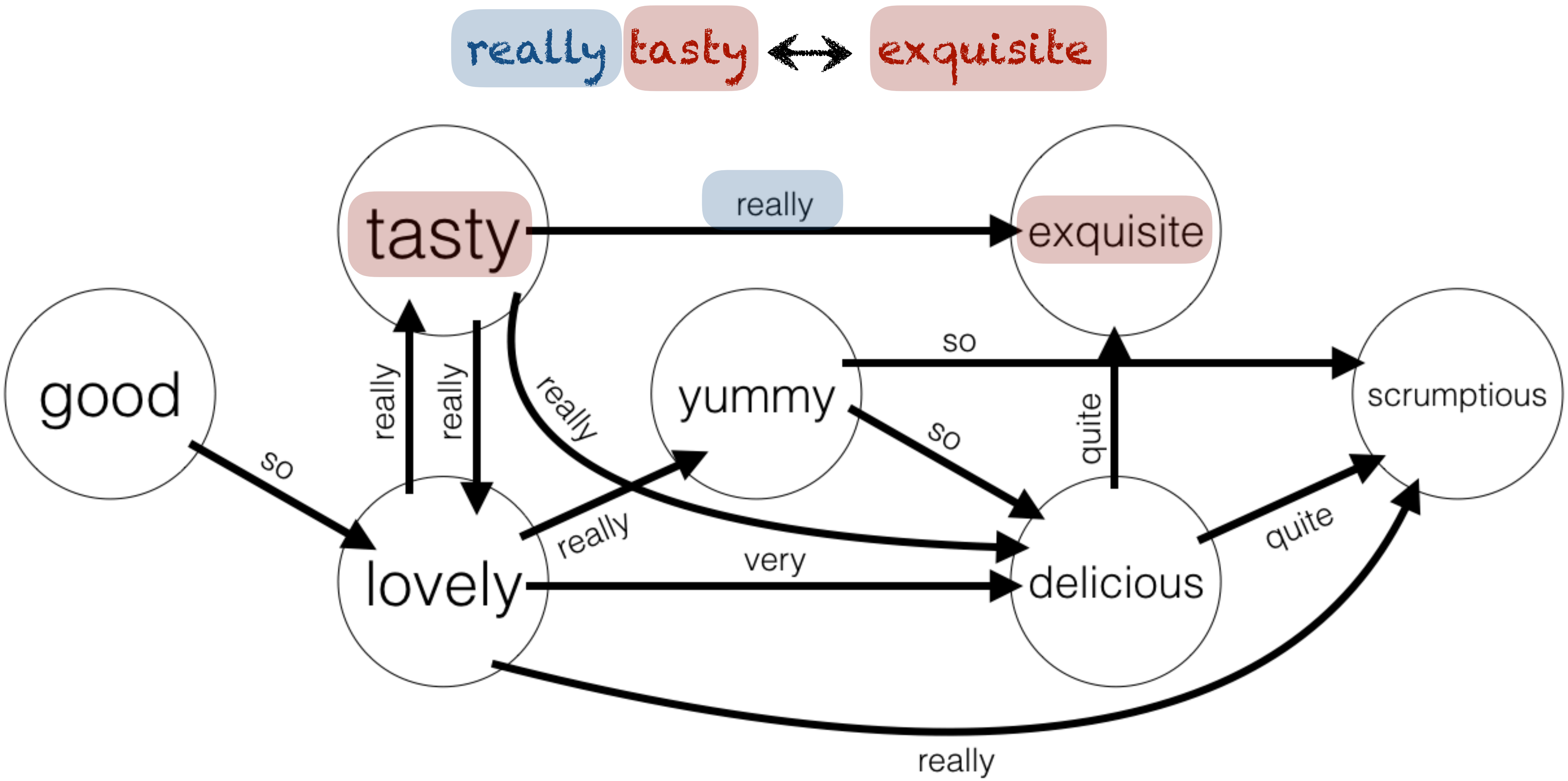
"JJGraph": 3,704 nodes (adjectives) and ~36k edges (610 unique adverb labels)

A graph structure can be used to encode the PPDB paraphrases matching our template



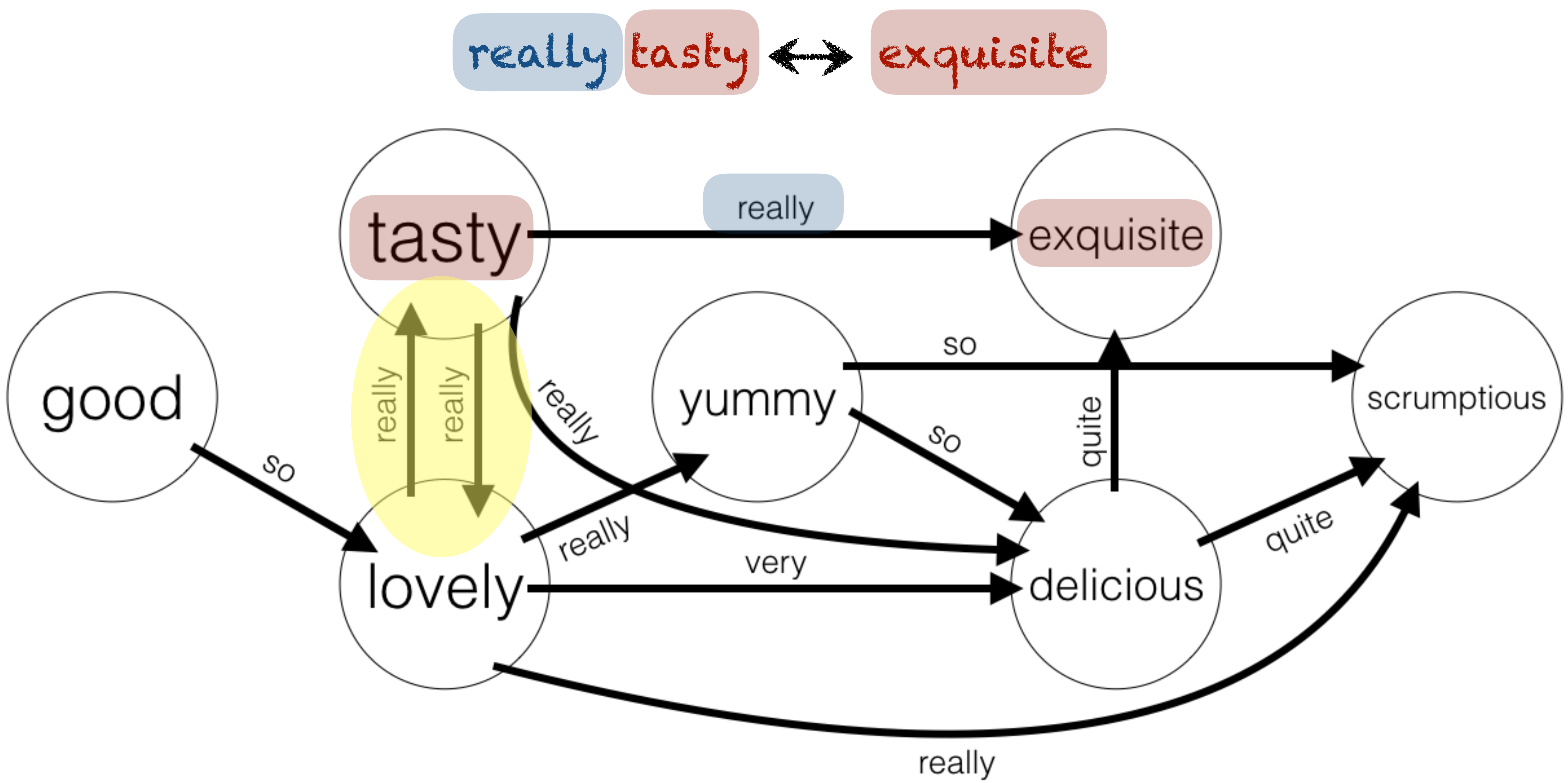
“`JJGraph`”: 3,704 nodes (adjectives) and ~36k edges (610 unique adverb labels)

A graph structure can be used to encode the PPDB paraphrases matching our template



"JJGraph": 3,704 nodes (adjectives) and ~36k edges (610 unique adverb labels)

A graph structure can be used to encode the PPDB paraphrases matching our template



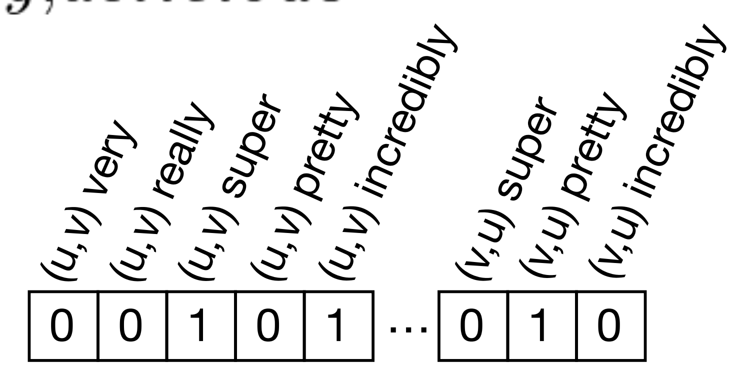
“JJGraph”: 3,704 nodes (adjectives) and ~36k edges (610 unique adverb labels)

Using paraphrase-based signals to predict relative adjective intensity

- Challenge 1: Identify intensifying adverbs
- Challenge 2: Resolve noise

x	y
(j_u, j_v)	1
(j_v, j_u)	0

$$x_{tasty,delicious} =$$



Using paraphrase-based signals to predict relative adjective intensity

- Challenge 1: Identify intensifying adverbs
- Challenge 2: Resolve noise
- Result: Relative intensity prediction model

$$score_{\text{pp}}(j_u, j_v) = \frac{1}{1 + \exp^{-wx_{uv}}} - 0.5$$

$$x_{tasty,delicious} = \begin{array}{cccccccc} \text{(u,v) very} & \text{(u,v) really} & \text{(u,v) super} & \text{(u,v) pretty} & \text{(u,v) incredibly} & & \text{(v,u) super} & \text{(v,u) pretty} & \text{(v,u) incredibly} \\ \boxed{0} & \boxed{0} & \boxed{1} & \boxed{0} & \boxed{1} & \cdots & \boxed{0} & \boxed{1} & \boxed{0} \end{array}$$

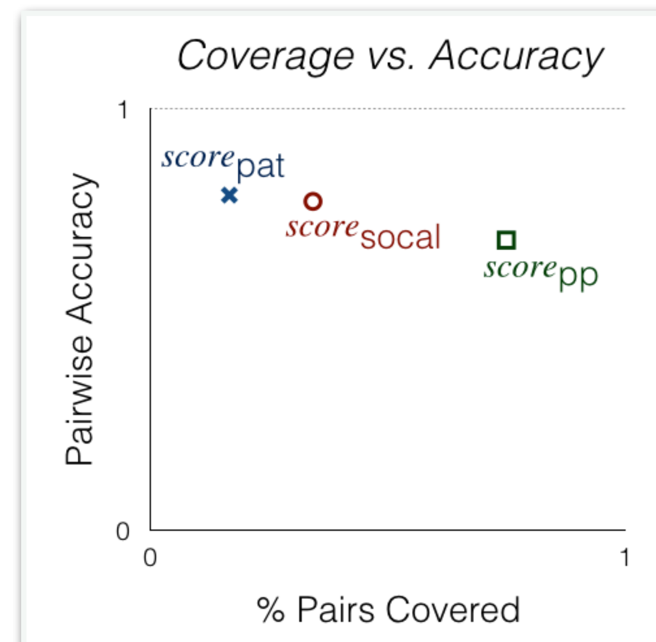
Learning Scalar Adjective Intensity

EMNLP 2018

hot < fiery

- Claims:
  - We can use adjectival phrase paraphrases to predict relative adjective intensity
  - This paraphrase-based information is complementary to pattern- and lexicon-based information

really hot <--> fiery



Other evidence types:
Lexicon-based evidence

Semantic Orientation CALculator (SOCAL)

Adjective	Score
exquisite	5
beautiful	4
appealing	3
above-average	2
okay	1
ho-hum	-1
pedestrian	-2
gross	-3
grisly	-4
abhorrent	-5

Taboada et al. 2011

Other evidence types: **Lexicon-based** evidence

Semantic Orientation CALculator (SOCAL)

Adjective	Score
exquisite	5
beautiful	4
appealing	3
above-average	2
okay	1
ho-hum	-1
pedestrian	-2
gross	-3
grisly	-4
abhorrent	-5

Taboada et al. 2011

+

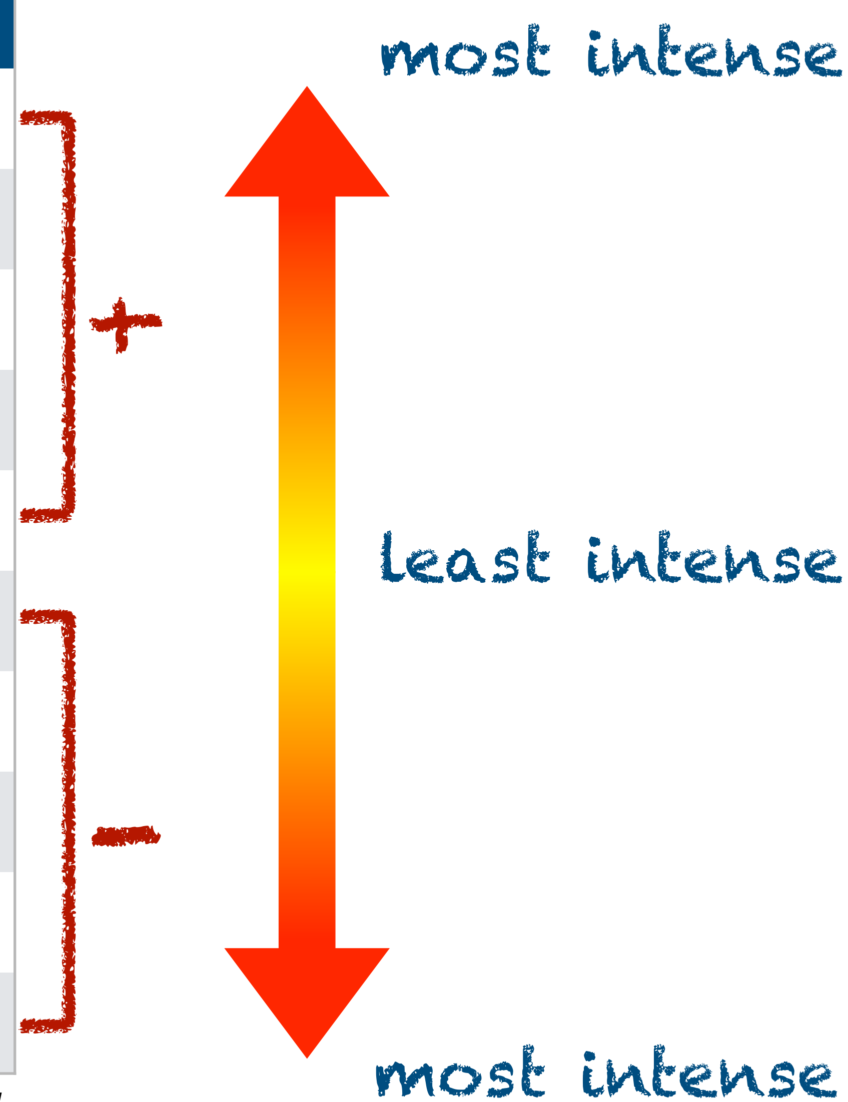
-

Other evidence types:
Lexicon-based evidence

Semantic Orientation CALculator
(SOCAL)

Adjective	Score
exquisite	5
beautiful	4
appealing	3
above-average	2
okay	1
ho-hum	-1
pedestrian	-2
gross	-3
grisly	-4
abhorrent	-5

Taboada et al. 2011



Other evidence types: Lexicon-based evidence

Semantic Orientation CALculator (SOCAL)

Adjective	Score
exquisite	5
beautiful	4
appealing	3
above-average	2
okay	1
ho-hum	-1
pedestrian	-2
gross	-3
grisly	-4
abhorrent	-5

Taboada et al. 2011

- Lexicon-based score simply requires a look-up in SOCAL
- In order to compute a score for (j_u, j_v), both adjectives must have the **same polarity**

$$score_{\text{socal}}(j_u, j_v) = |L(j_v)| - |L(j_u)|, \quad \textbf{iff sign}(j_u) = \textbf{sign}(j_v)$$

Other evidence types:
Pattern-based evidence

"The show was **funny**, but not **hilarious**." → funny < hilarious

"It's not **freezing**, but still **cold**." → cold < freezing

Other evidence types: Pattern-based evidence

"The show was **funny**, but not **hilarious**." → funny < hilarious

"It's not **freezing**, but still **cold**." → cold < freezing

- We use DeMelo & Bansal ('13) method for producing a pattern-based score
  - Extract *weak-strong* (W) and *strong-weak* (S) patterns from Google *n*-gram corpus

Other evidence types: Pattern-based evidence

"The show was **funny**, but not **hilarious**." → funny < hilarious

"It's not **freezing**, but still **cold**." → cold < freezing

- We use DeMelo & Bansal ('13) method for producing a pattern-based score

Other evidence types: Pattern-based evidence

"The show was **funny**, but not **hilarious**." → funny < hilarious

"It's not **freezing**, but still **cold**." → cold < freezing

- We use DeMelo & Bansal ('13) method for producing a pattern-based score

$$score_{\text{pat}}(j_u, j_v) = \frac{(W_u - S_u) - (W_v - S_v)}{\text{count}(j_u) \cdot \text{count}(j_v)}$$

Other evidence types: Pattern-based evidence

"The show was **funny**, but not **hilarious**." ⟶ funny < hilarious

"It's not **freezing**, but still **cold**." ⟶ cold < freezing

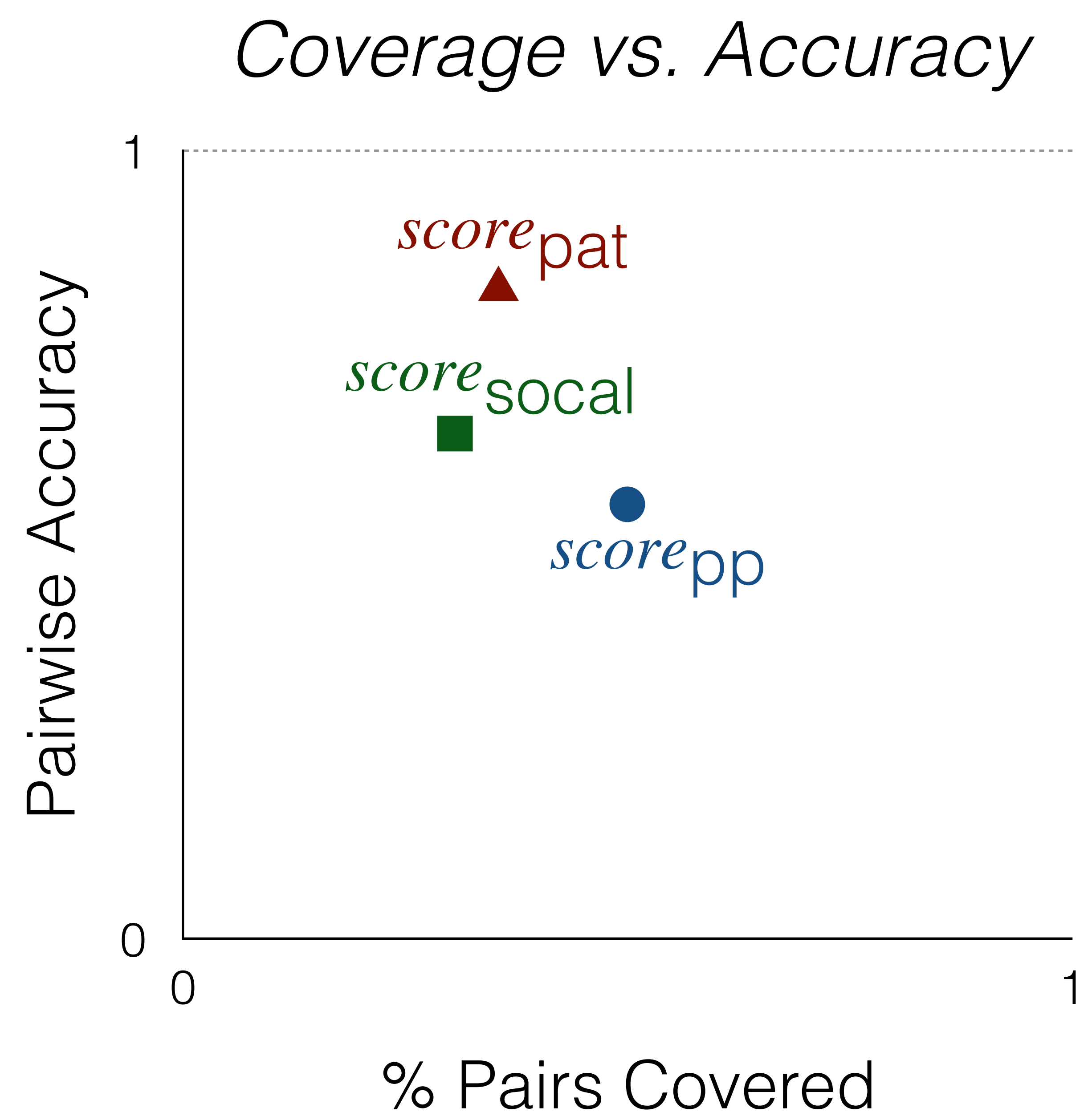
- We use DeMelo & Bansal ('13) method for producing a pattern-based score

$$score_{\text{pat}}(j_u, j_v) = \frac{\overbrace{(W_u - S_u)}^{\text{evidence of } j_u < j_v} - \overbrace{(W_v - S_v)}^{\text{evidence of } j_v < j_u}}{\text{count}(j_u) \cdot \text{count}(j_v)}$$

Paraphrase evidence has high coverage, but other types are more accurate

- For each score type, predict intensity direction for adjective pairs from 3 datasets (878 pairs total)

- Report % pairs covered, and directional accuracy



We can combine score types using a back-off method

$$score_{\mathrm{x+y}}(j_u, j_v)$$

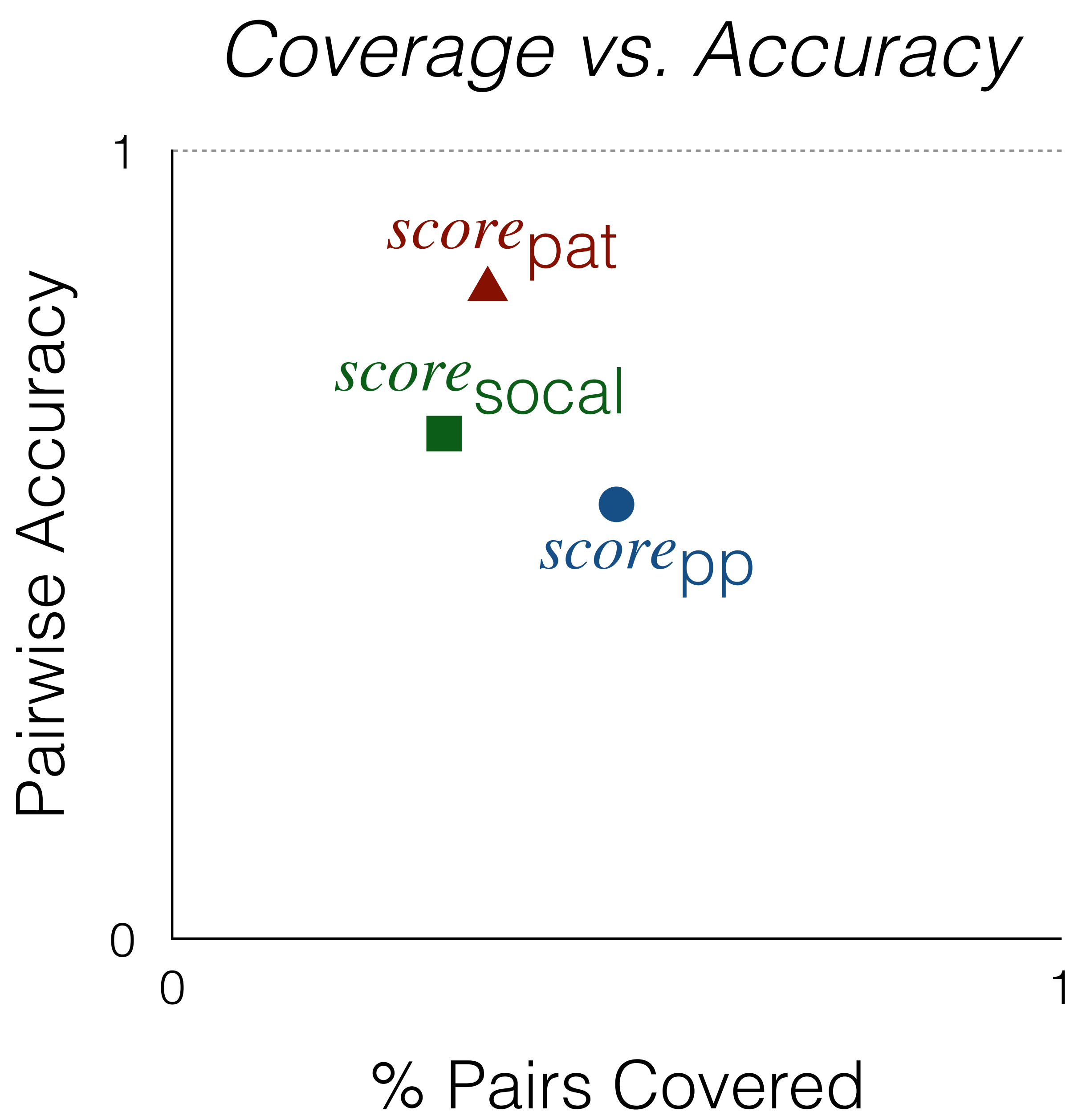
We can combine score types using a back-off method

$$score_{\mathrm{x+y}}(j_u, j_v)$$

"If $score_x$ can be computed, use it. Otherwise, use $score_y$."

Combining complementary score types yields higher coverage and accuracy

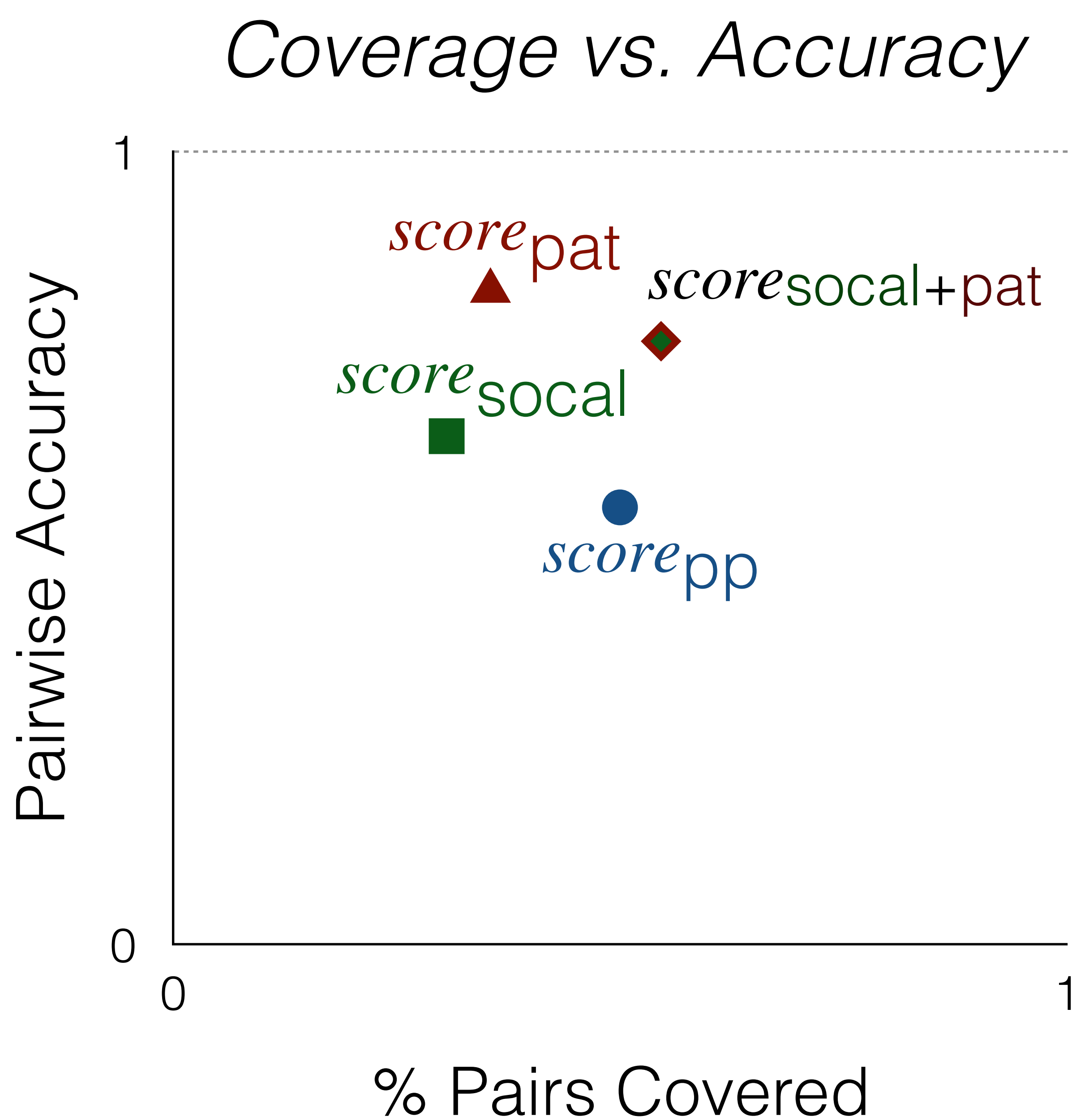
- For each score type, predict intensity direction for adjective pairs from 3 datasets (878 pairs total)
- Report % pairs covered, and directional accuracy



Combining complementary score types yields higher coverage and accuracy

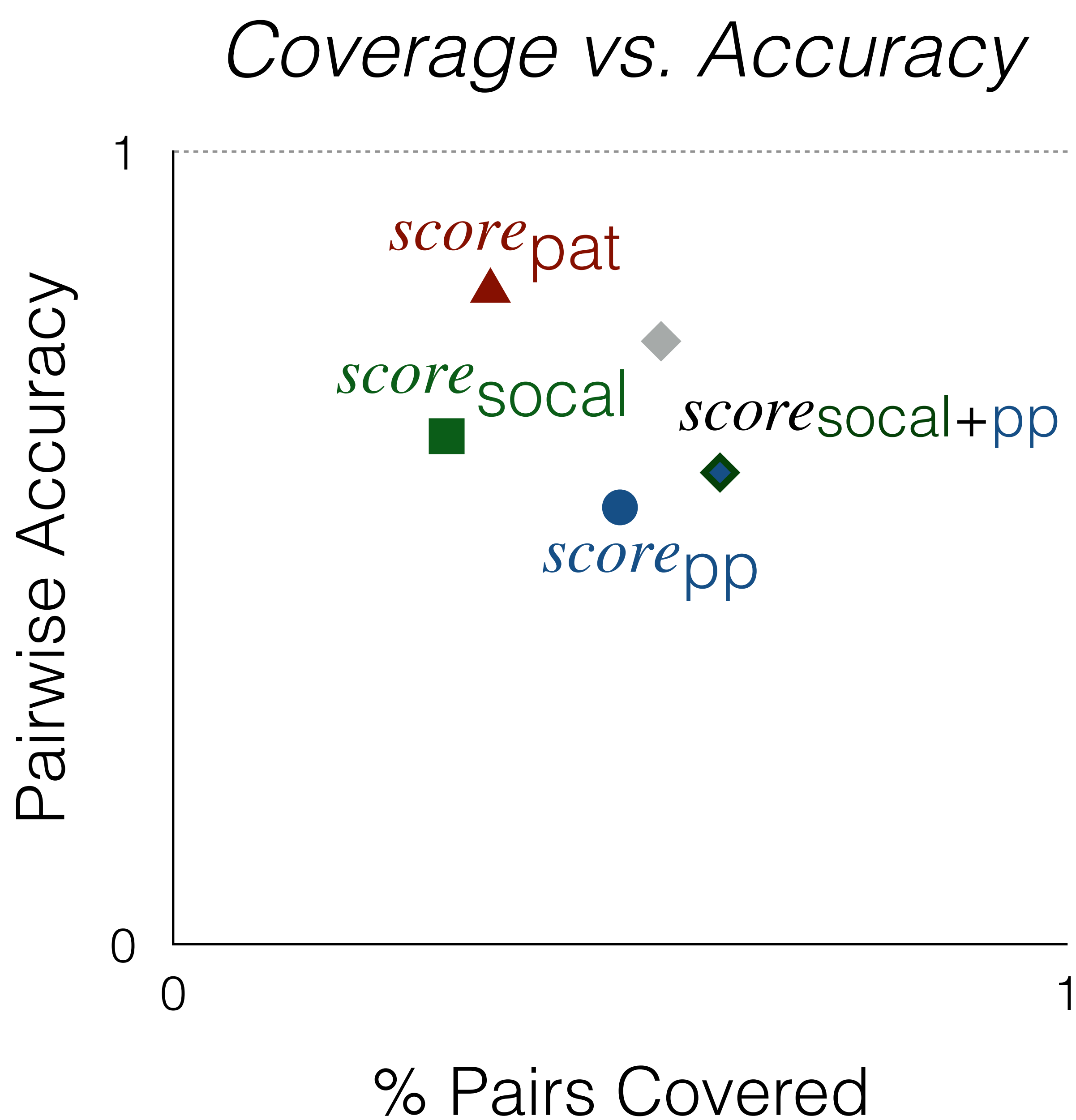
- For each score type, predict intensity direction for adjective pairs from 3 datasets (878 pairs total)

- Report % pairs covered, and directional accuracy



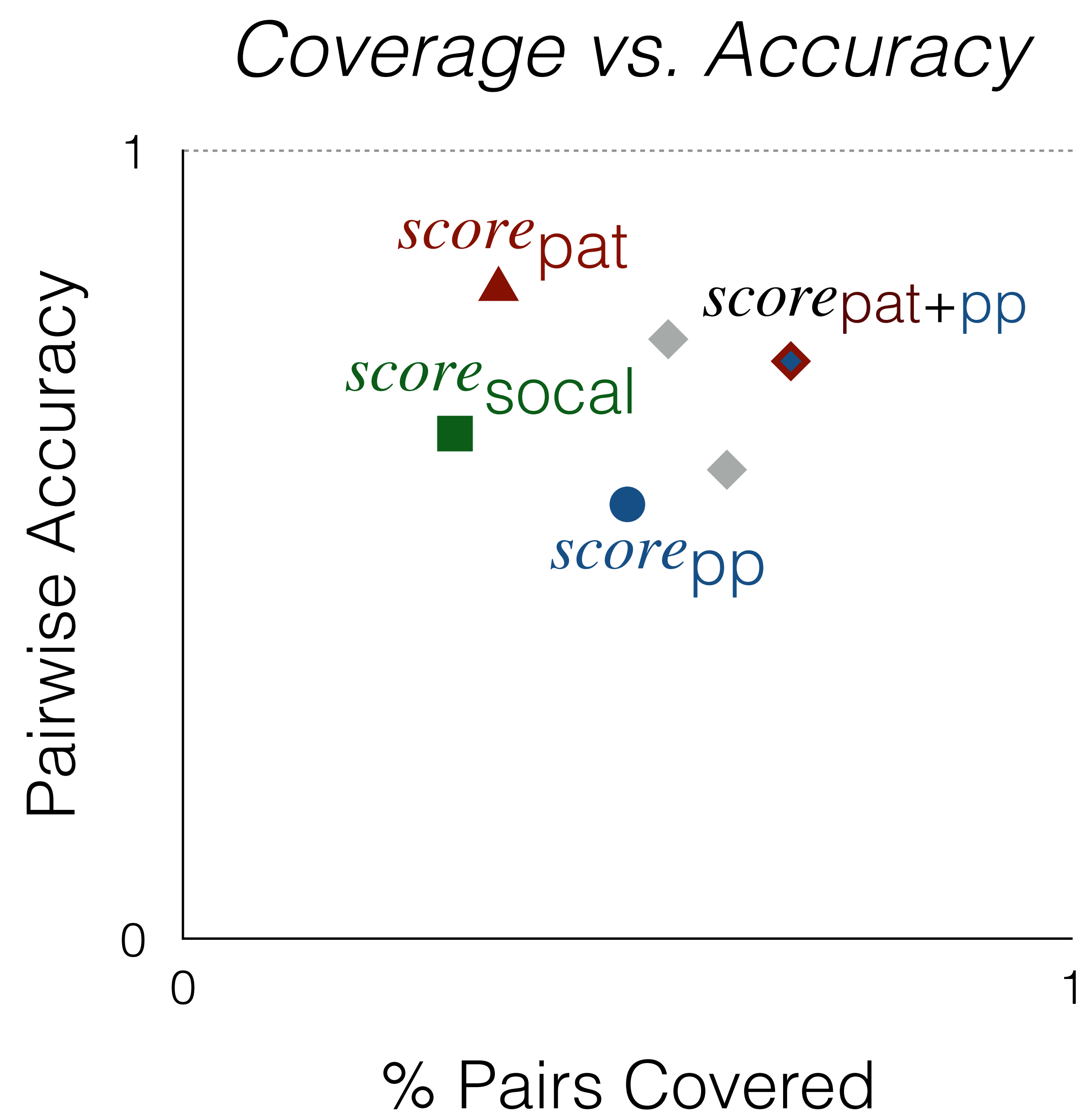
Combining complementary score types yields higher coverage and accuracy

- For each score type, predict intensity direction for adjective pairs from 3 datasets (878 pairs total)
- Report % pairs covered, and directional accuracy



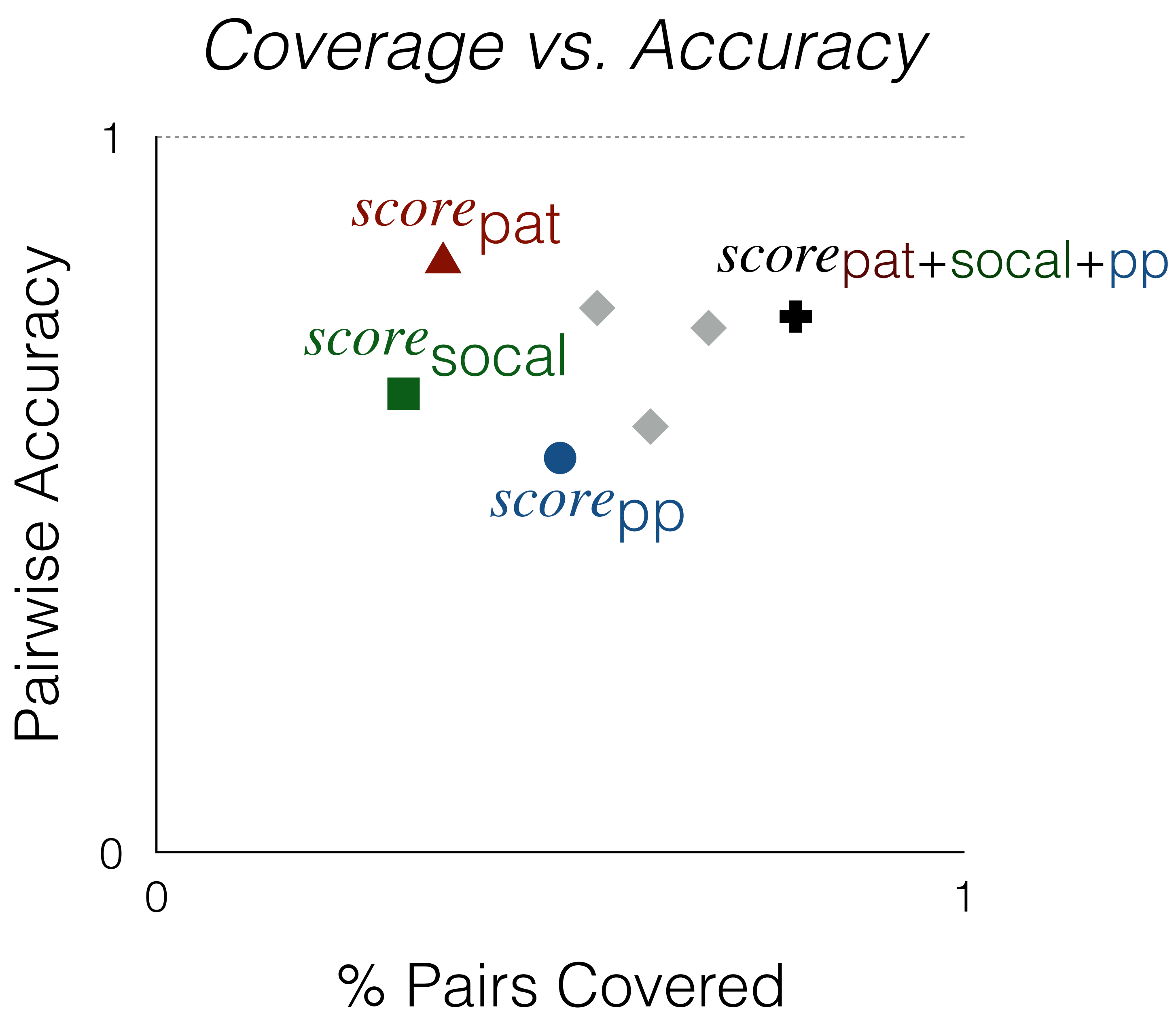
Combining complementary score types yields higher coverage and accuracy

- For each score type, predict intensity direction for adjective pairs from 3 datasets (878 pairs total)
- Report % pairs covered, and directional accuracy



Combining complementary score types yields higher coverage and accuracy

- For each score type, predict intensity direction for adjective pairs from 3 datasets (878 pairs total)
- Report % pairs covered, and directional accuracy



Experimental setup: Indirect Question Answering

- IDQA Dataset (deMarneffe et al. 2010)
  - 123 question/answer pairs
- Rule-based method for predicting the answer (deMarneffe et al. 2010)

Experimental setup: Indirect Question Answering

Q: *Was he a successful ruler?*

A: *Oh, a tremendous ruler.*

(YES!)

Q: *Does it have a large impact?*

A: *It has a medium-sized impact.*

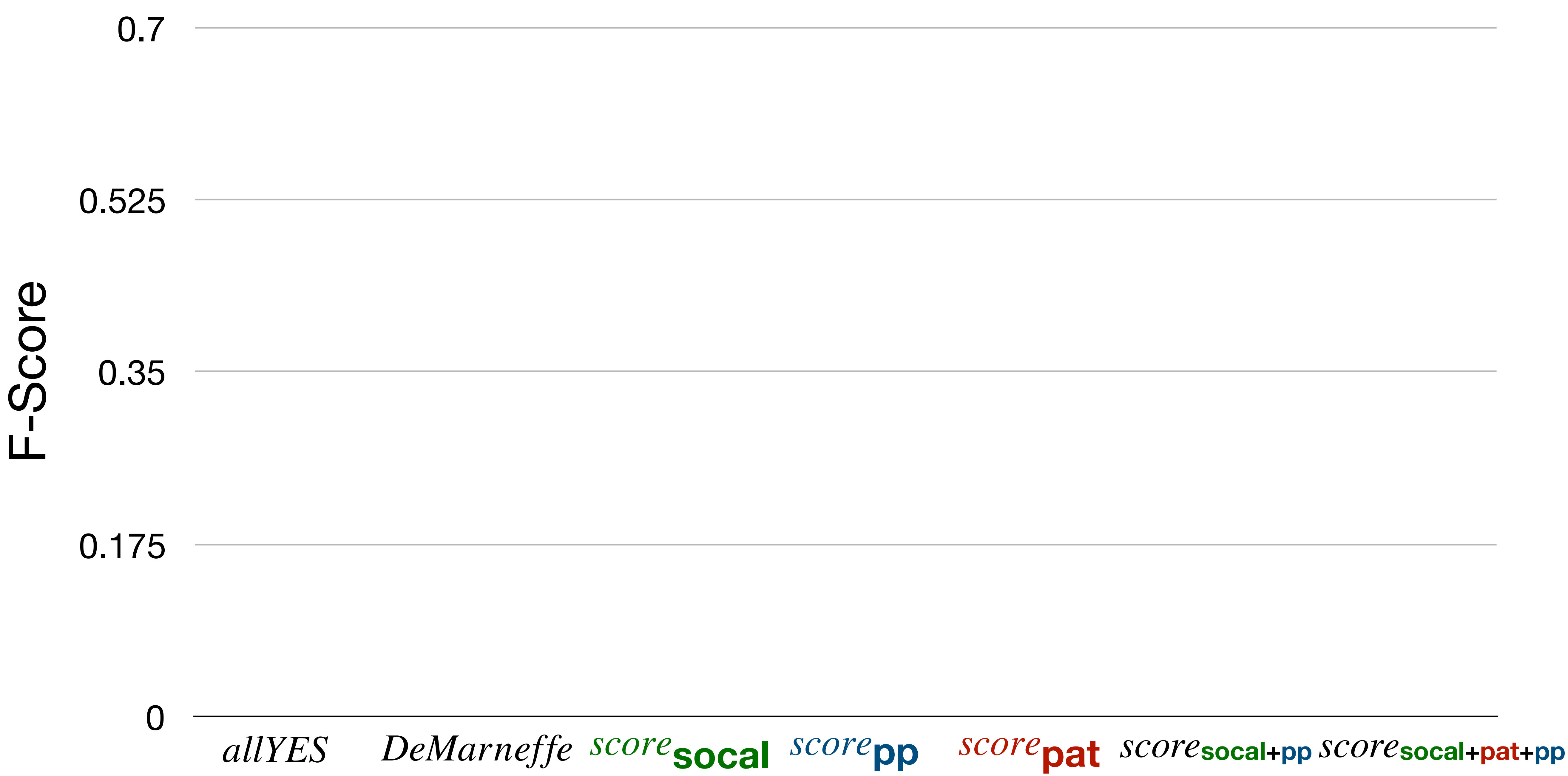
(NO!)

- IDQA Dataset (deMarneffe et al. 2010)
  - 123 question/answer pairs
- Rule-based method for predicting the answer (deMarneffe et al. 2010)

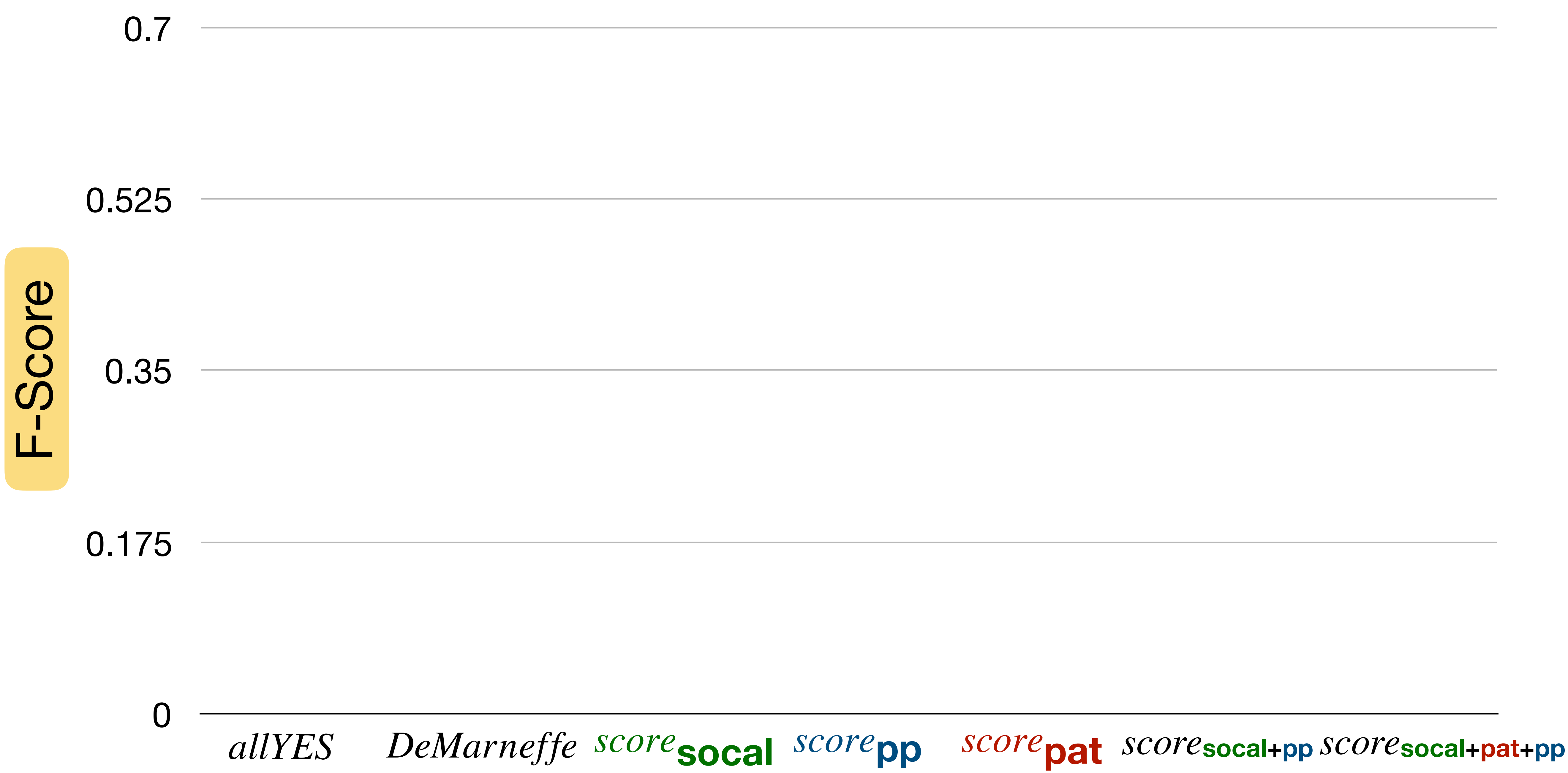
Again, combining paraphrase with other types of evidence leads to strongest overall results

allYES *DeMarneffe* $score_{\mathbf{socal}}$ $score_{\mathbf{pp}}$ $score_{\mathbf{pat}}$ $score_{\mathbf{socal+pp}}$ $score_{\mathbf{socal+pat+pp}}$

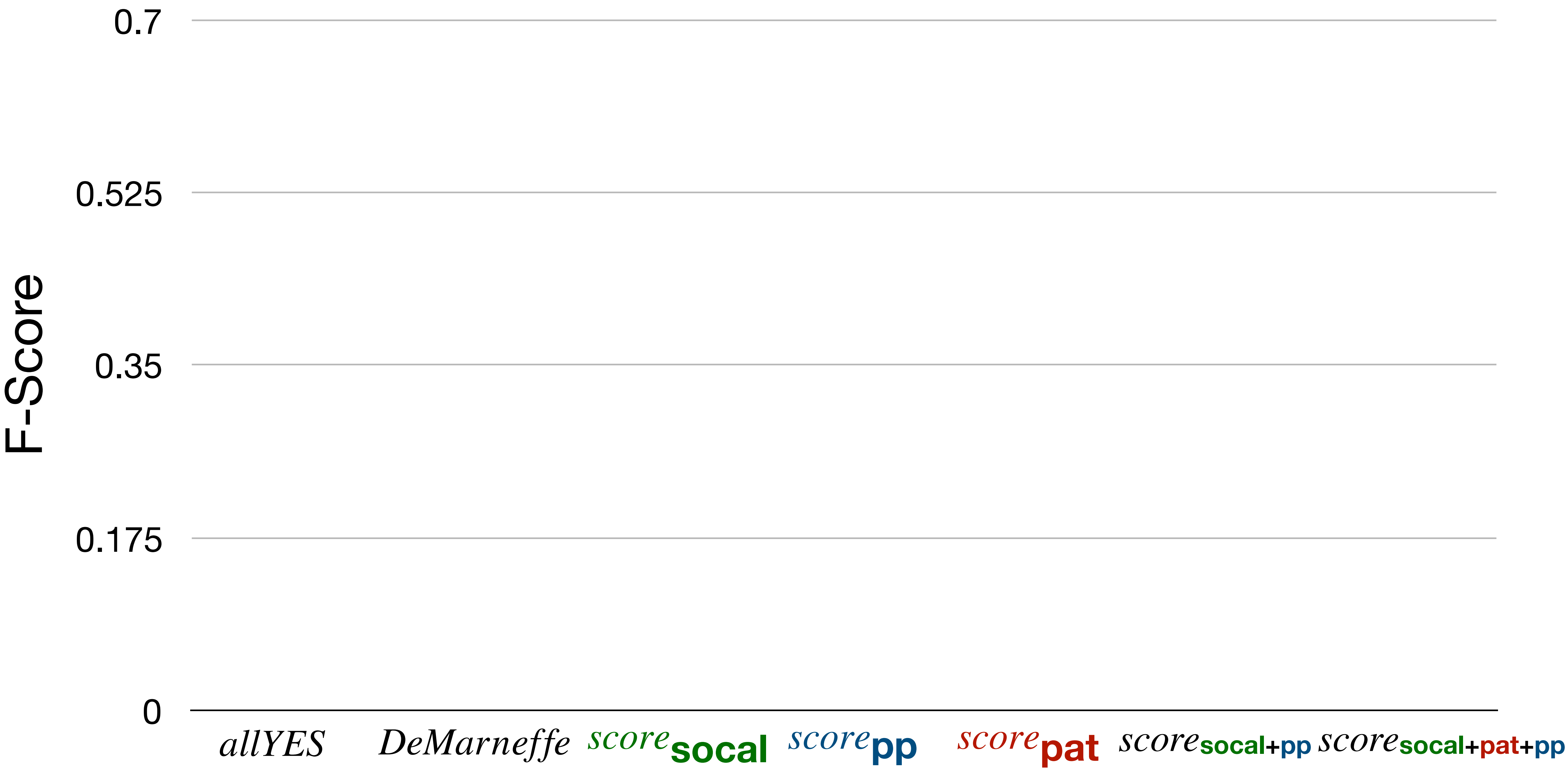
Again, combining paraphrase with other types of evidence leads to strongest overall results



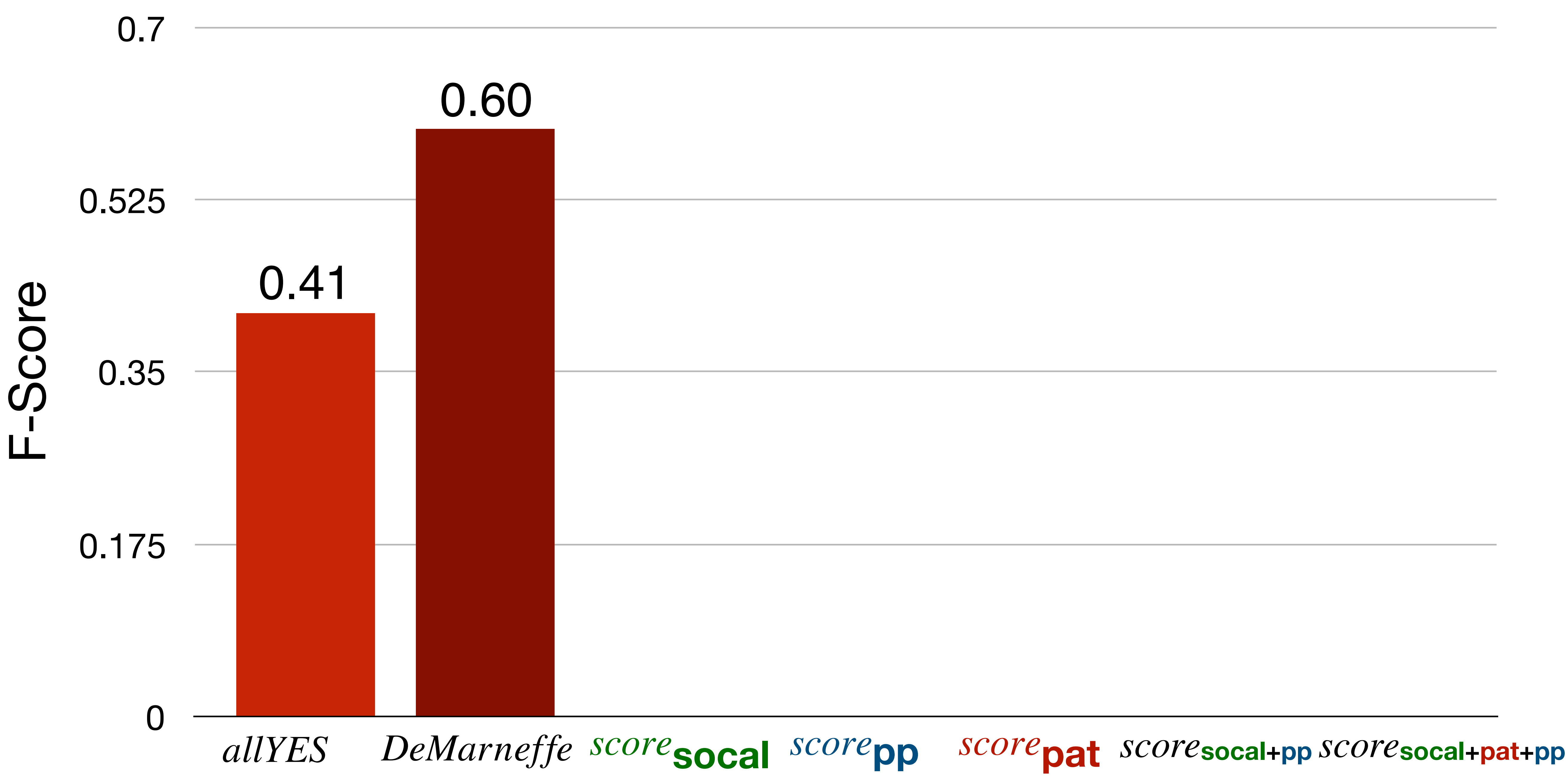
Again, combining paraphrase with other types of evidence leads to strongest overall results



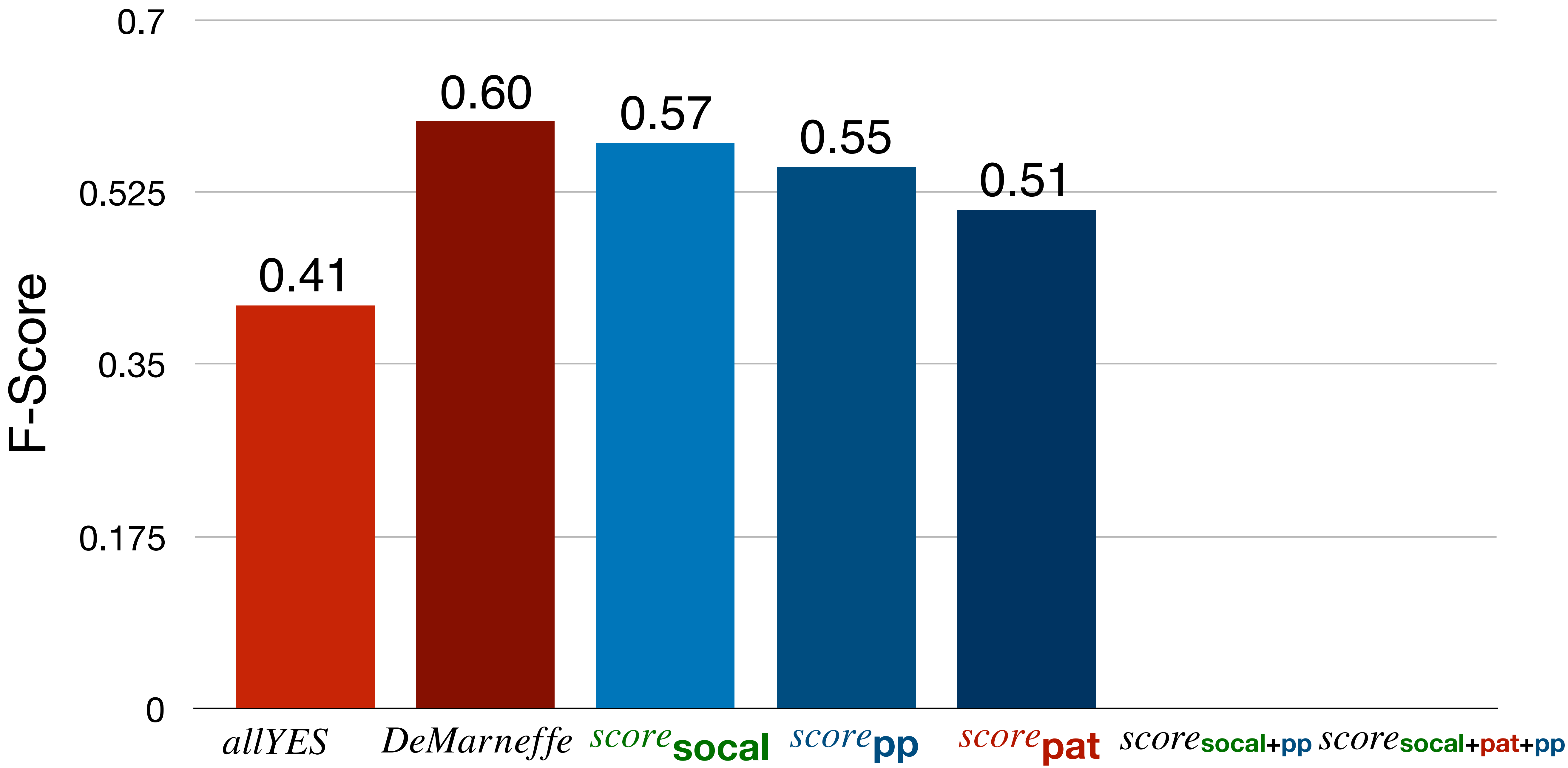
Again, combining paraphrase with other types of evidence leads to strongest overall results



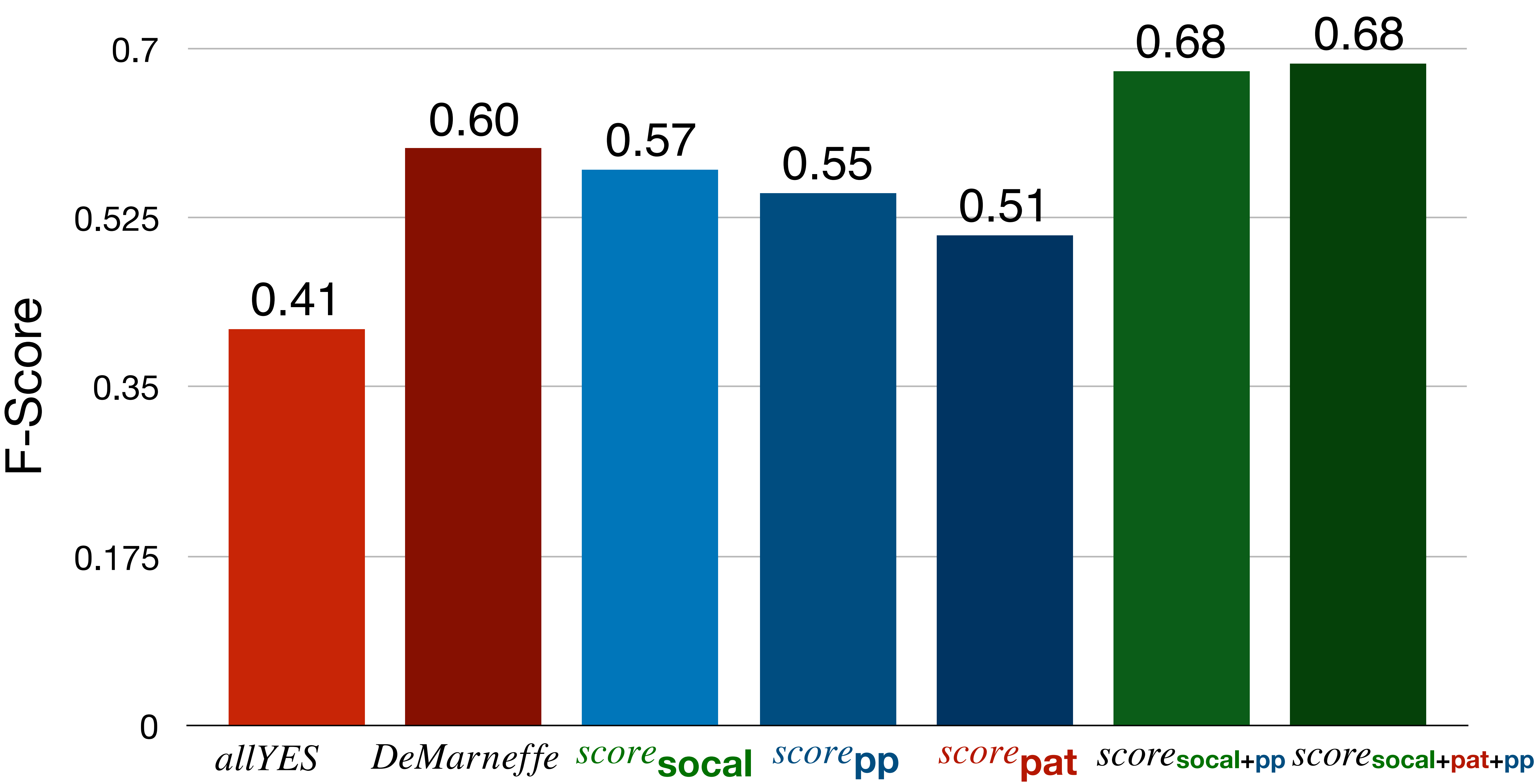
Again, combining paraphrase with other types of evidence leads to strongest overall results



Again, combining paraphrase with other types of evidence leads to strongest overall results



Again, combining paraphrase with other types of evidence leads to strongest overall results



Learning Scalar Adjective Intensity

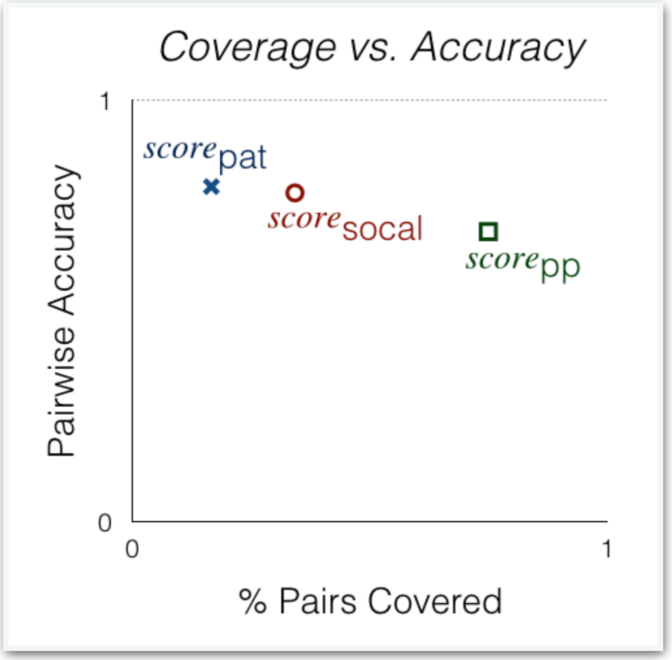
EMNLP 2018

hot < fiery

- Claims:
  - We can use adjectival phrase paraphrases to predict relative adjective intensity

really hot <--> fiery

  - This paraphrase-based information is complementary to pattern- and lexicon-based information



Learning Scalar Adjective Intensity

EMNLP 2018

hot < fiery

Learning Scalar Adjective Intensity

EMNLP 2018

hot < fiery

- Take-aways:

Learning Scalar Adjective Intensity

EMNLP 2018

hot < fiery

- Take-aways:
  - Paraphrases provide a new method for predicting relative adjective intensity

Learning Scalar Adjective Intensity

EMNLP 2018

hot < fiery

- Take-aways:
    - Paraphrases provide a new method for predicting relative adjective intensity
    - With higher coverage and lower precision, paraphrase-based intensity evidence is complementary to lexicon- and pattern-based intensity evidence

Motivation

Using Paraphrases to Model Word Sense

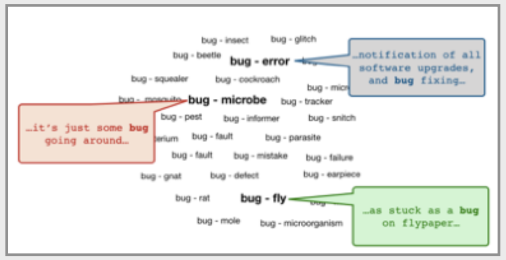
NAACL 2016; SENSE@EACL 2017

Learning Scalar Adjective Intensity

EMNLP 2018

Meaning-specific Examples of Word Use

In submission



Conclusion

Word meaning is contextual.

Word meaning is contextual.

Premise	`He rearranged the layout of the room, placing the table by the window.`
Hypothesis	`The furniture was moved.`
Entailed?	

Word meaning is contextual.

Premise	He **rearranged** the layout of the room, placing the **table** by the window.
Hypothesis	The **furniture** **was moved**.
Entailed?	**TRUE**

Word meaning is contextual.

Premise	`He rearranged the layout of the room, placing the table by the window.`
Hypothesis	`The furniture was moved.`
Entailed?	**`TRUE`**

Premise	`She rearranged the layout of the document, placing the table on page four.`
Hypothesis	`The furniture was moved.`
Entailed?	

Word meaning is contextual.

Premise	He rearranged the layout of the room, placing the table by the window.
Hypothesis	The furniture was moved.
Entailed?	**TRUE**

Premise	She **rearranged** the layout of the document, placing the **table** on page four.
Hypothesis	The **furniture** **was moved**.
Entailed?	**FALSE**

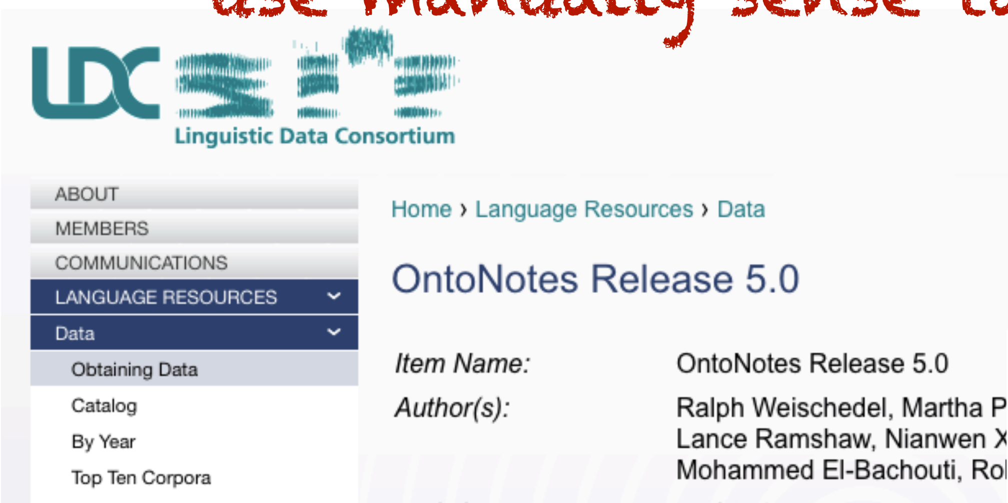
How can we create corpora that explicitly model different meanings?

How can we create corpora that explicitly model different meanings?

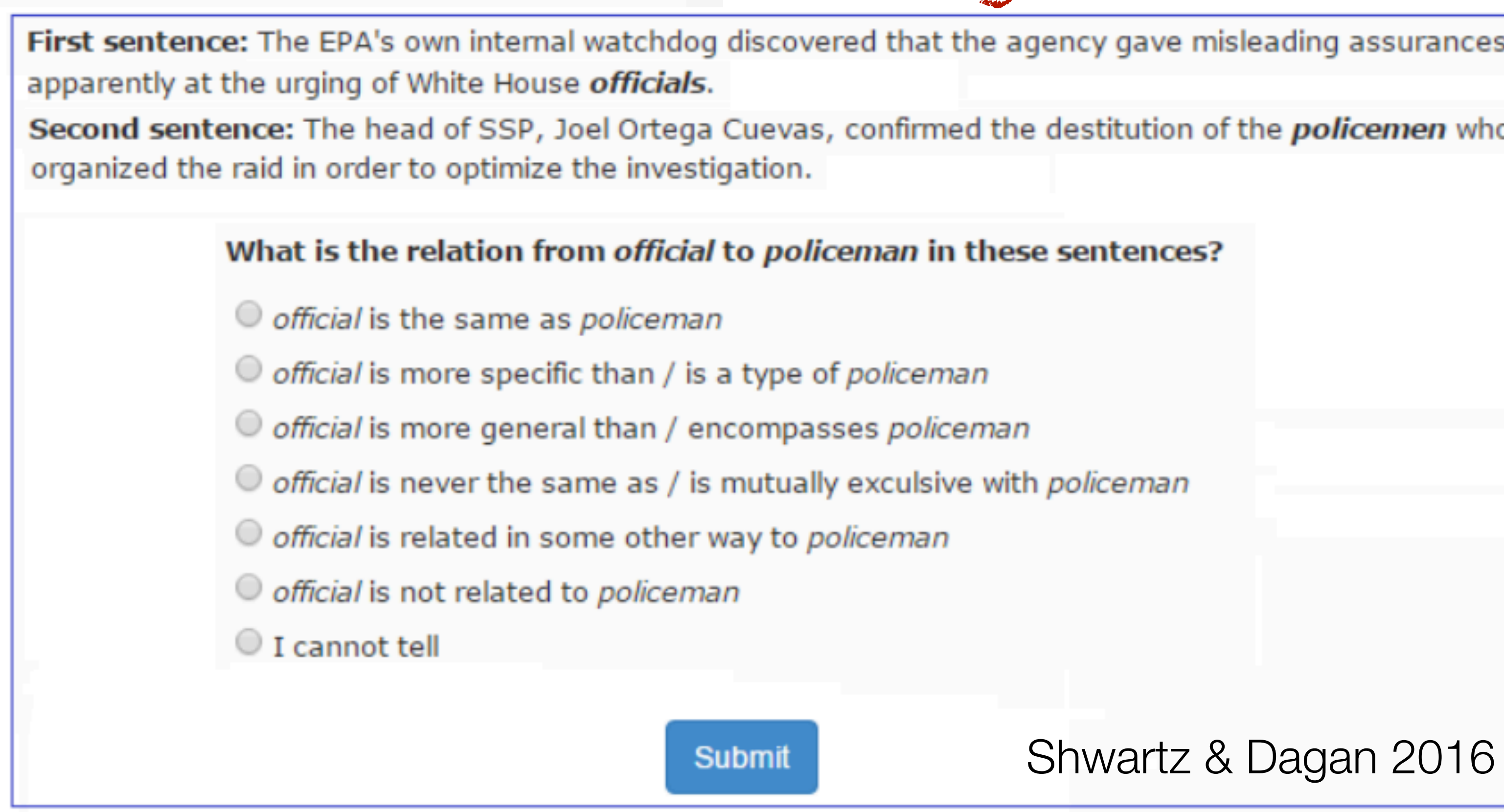
use manually sense-tagged resources

How can we create corpora that explicitly model different meanings?

use manually sense-tagged resources



crowdsourcing

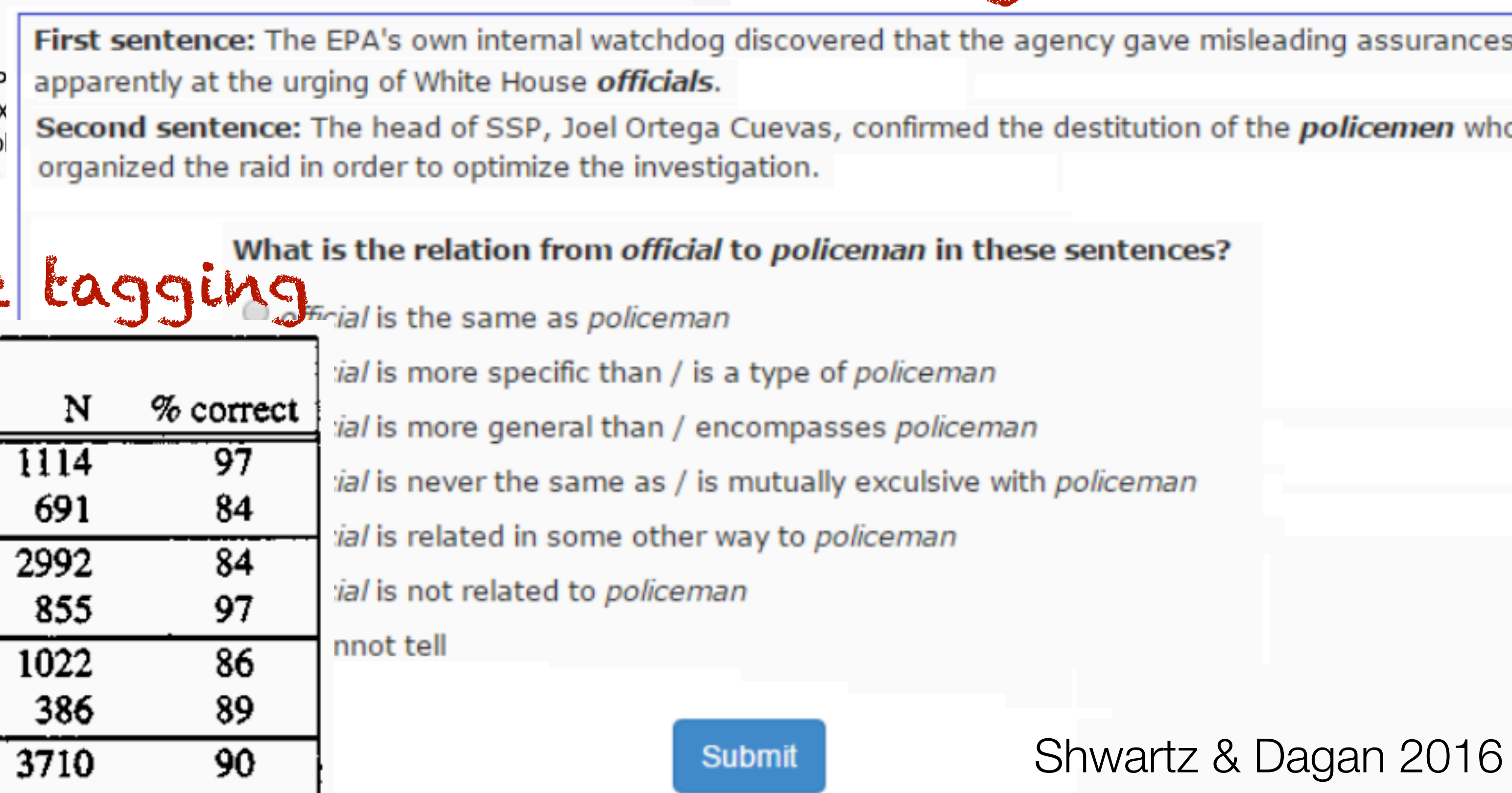


Shwartz & Dagan 2016

How can we create corpora that explicitly model different meanings?

use manually sense-tagged resources

crowdsourcing



Shwartz & Dagan 2016

unsupervised sense tagging

Table 3: Six Polysemous Words

English	French	sense	N	% correct
duty	droit	tax	1114	97
	devoir	obligation	691	84
drug	médicament	medical	2992	84
	drogue	illicit	855	97
land	terre	property	1022	86
	pays	country	386	89
language	langue	medium	3710	90
	langage	style	170	91
position	position	place	5177	82
	poste	job	577	86
sentence	peine	judicial	296	97
	phrase	grammatical	148	100

Gale et al. 1992

Paraphrase Sense-Tagged Sentences (PSTS)

bug - insect
bug - glitch
bug - beetle
bug - error
bug - squealer
bug - cockroach
bug - micr
bug - mosquito
bug - microbe
bug - tracker
bug - pest
bug - informer
bug - snitch
terium
bug - fault
bug - parasite
bug - fault
bug - mistake
bug - failure
bug - gnat
bug - defect
bug - earpiece
bug - rat
bug - fly
bug - mole
bug - microorganism

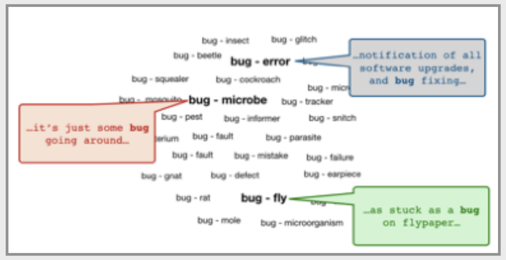
…notification of all software upgrades, and **bug** fixing…

…it's just some **bug** going around…

…as stuck as a **bug** on flypaper…

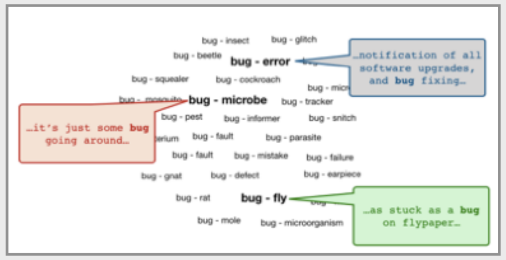
Meaning-specific Examples of Word Use

In submission



Meaning-specific Examples of Word Use

In submission



- Claims:
  - The pivot method can be applied to generate a paraphrase-sense-tagged corpus at scale
  - The resulting resource is useful for training sense-aware models for downstream tasks

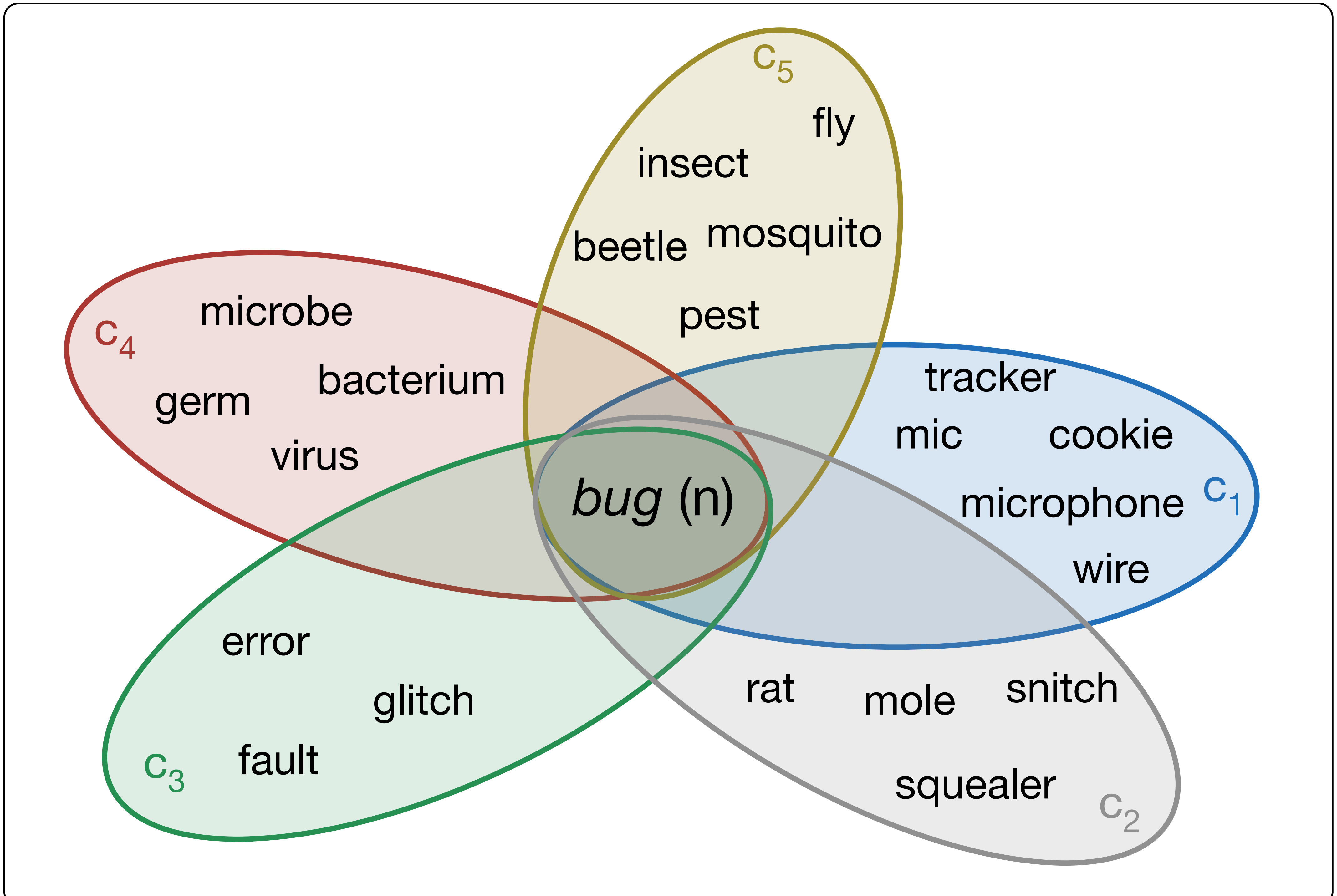
Paraphrases and Polysemy

Each paraphrase of a target word represents a slightly distinct meaning

fly
insect
microbe
error
pest
bacterium
beetle
snitch
tracker
mole
cookie
bug (n)
mic
germ
microphone
virus
wire
rat
glitch
fault
mosquito
squealer

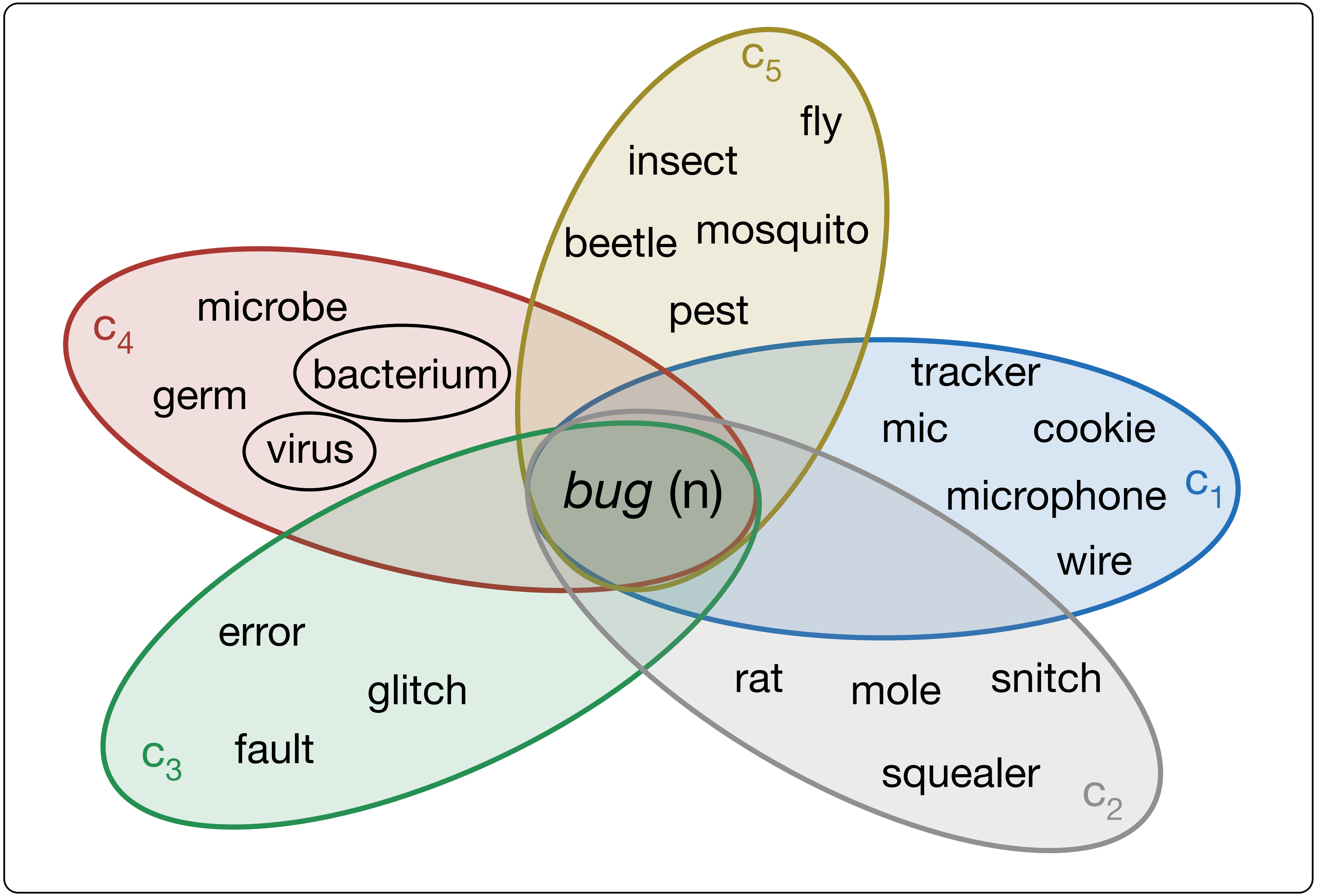
Paraphrases and Polysemy

Each paraphrase of a target word represents a slightly distinct meaning



Paraphrases and Polysemy

Each paraphrase of a target word represents a slightly distinct meaning



Paraphrase Sense-Tagged Sentences (PSTS)

bug - insect

bug - glitch

bug - beetle

bug - error

…notification of all software upgrades, and **bug** fixing…

bug - squealer

bug - cockroach

bug - micr

bug - mosquito

bug - microbe

bug - tracker

…it's just some **bug** going around…

bug - pest

bug - informer

bug - snitch

terium

bug - fault

bug - parasite

bug - fault

bug - mistake

bug - failure

bug - gnat

bug - defect

bug - earpiece

bug - rat

bug - fly

…as stuck as a **bug** on flypaper…

bug - mole

bug - microorganism

Bilingual pivoting can be leveraged to build PSTS

Bilingual pivoting can be leveraged to build PSTS

… the nationalist bug has infected the EU itself …

le virus nationaliste a infecté jusqu'à l'union …

Bilingual pivoting can be leveraged to build PSTS

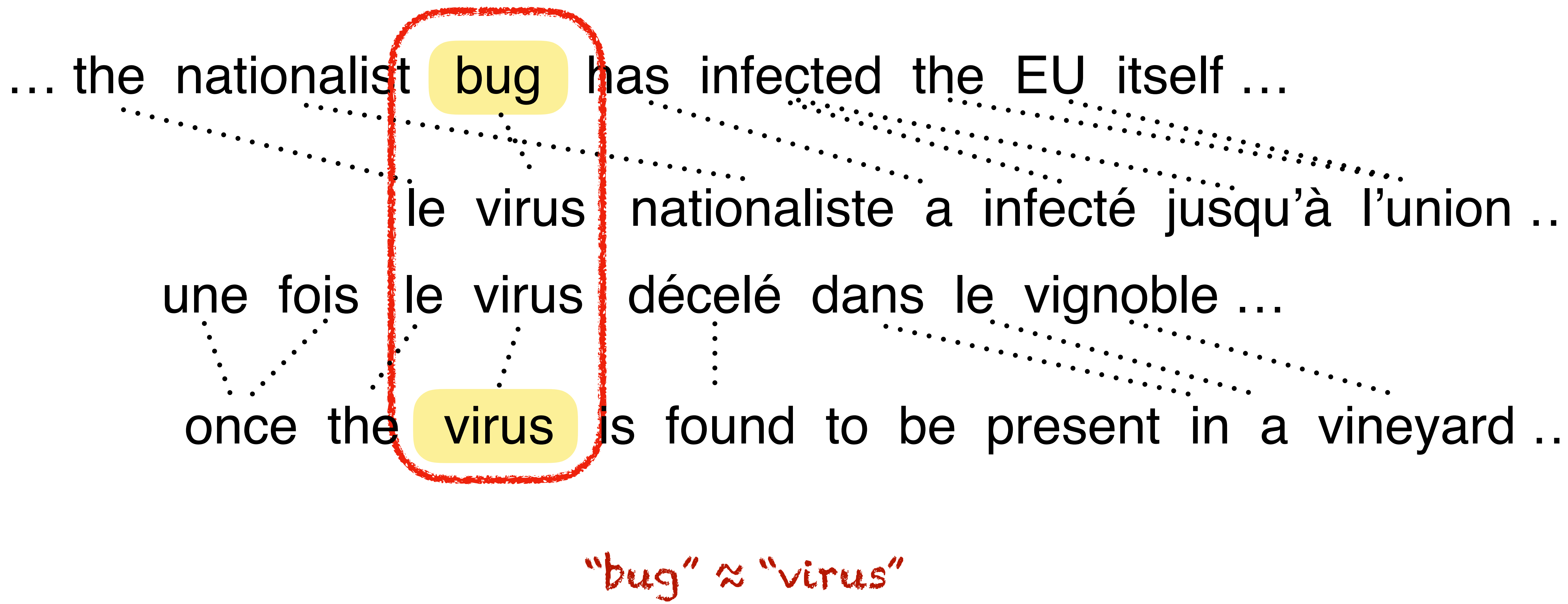
… the nationalist **bug** has infected the EU itself …

le virus nationaliste a infecté jusqu'à l'union …

une fois le virus décelé dans le vignoble …

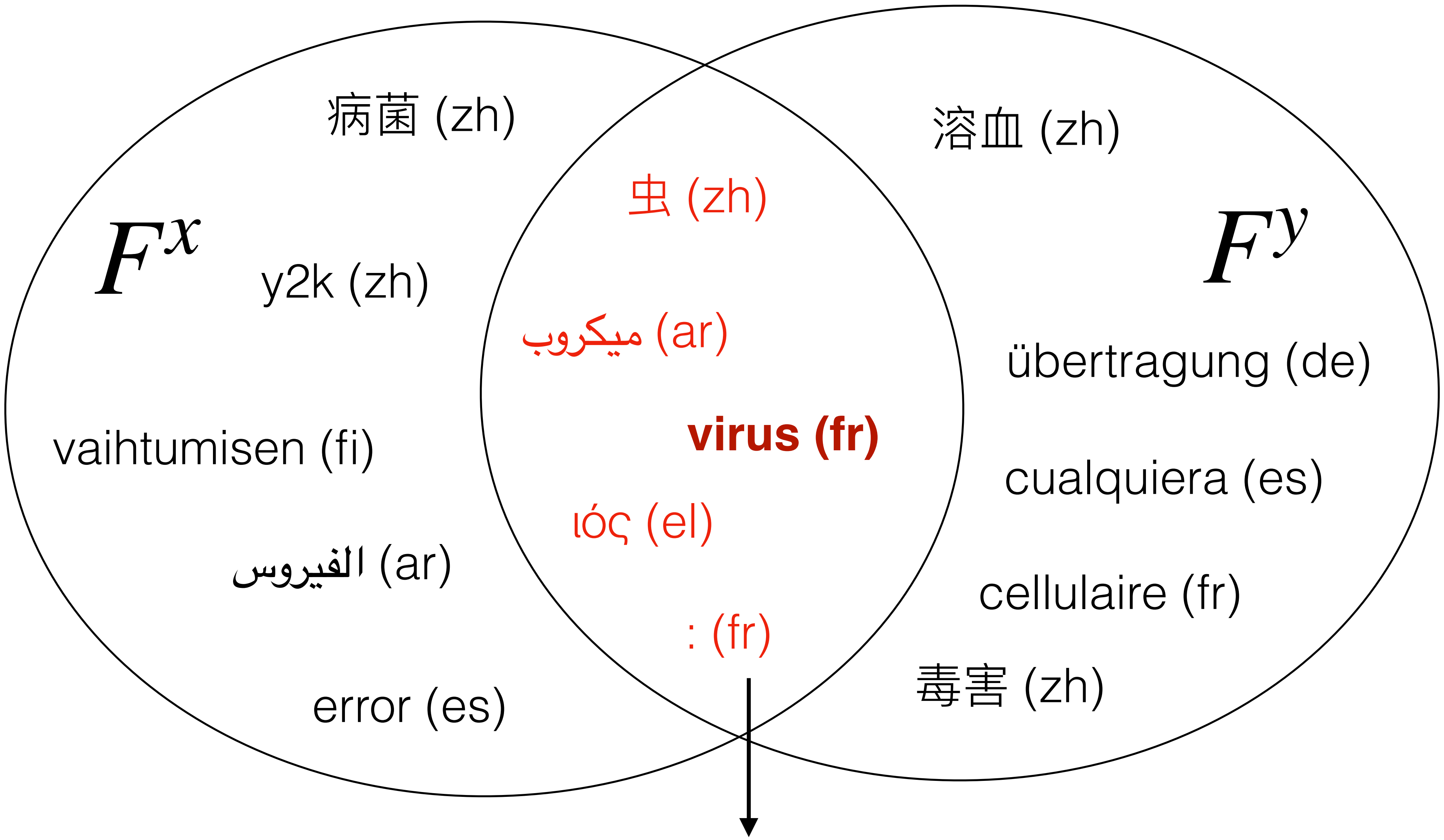
once the **virus** is found to be present in a vineyard …

Bilingual pivoting can be leveraged to build PSTS



Step 1: Find shared translations

x = *bug* y = *virus*



$$F^{xy} = F^x \cap F^y$$

Step 2: Prioritize Translations

x = *bug* y = *virus*

$$f \in F^{xy}$$

ιός (el) [virus]

virus (fr) [virus]

ميكروب (ar) [microbial]

虫 (zh) [worm]

: (fr) [<punctuation>]

Step 2: Prioritize Translations

x = *bug* y = *virus*

$f \in F^{xy}$
ιός (el) [virus]
virus (fr) [virus]
ميكروب (ar) [microbial]
虫 (zh) [worm]
: (fr) [<punctuation>]

$$PMI(y, f) = \frac{p(f \mid y)}{p(f)}$$

Step 2: Prioritize Translations

x = *bug*   y = *virus*

$f \in F^{xy}$	$\downarrow PMI(y, f)$
ιός (el) [virus]	11.4
virus (fr) [virus]	10.0
ميكروب (ar) [microbial]	6.5
虫 (zh) [worm]	3.4
: (fr) [<punctuation>]	-0.7

Step 3: Enumerate Sentences

x = *bug* y = *virus*

$f \in F^{xy}$	$\downarrow PMI(y, f)$
ιός (el) [virus]	11.4
virus (fr) [virus]	10.0
ميكروب (ar) [microbial]	6.5
虫 (zh) [worm]	3.4
: (fr) [<punctuation>]	-0.7

Step 3: Enumerate Sentences

x = *bug* y = *virus*

$f \in F^{xy}$	$\downarrow PMI(y, f)$
ιός (el) [virus]	11.4
virus (fr) [virus]	10.0
ميكروب (ar) [microbial]	6.5
虫 (zh) [worm]	3.4
: (fr) [<punctuation>]	-0.7

το 1999 , όλοι πίστευαν ότι ο **ιός** της χιλιετίας θα προκαλούσε παγκόσμια καταστροφή επηρεάζοντας όλα τα συστήματα υπολογιστών στον κόσμο .

In 1999, everybody believed that the millennium **bug** would create a global disaster by closing down computer systems across the world.

Step 3: Enumerate Sentences

x = *bug*   y = *virus*

$f \in F^{xy}$	$\downarrow PMI(y, f)$
ιός (el) [virus]	11.4
virus (fr) [virus]	10.0
ميكروب (ar) [microbial]	6.5
虫 (zh) [worm]	3.4
: (fr) [<punctuation>]	-0.7

το 1999 , όλοι πίστευαν ότι ο **ιός** της χιλιετίας θα προκαλούσε παγκόσμια καταστροφή επηρεάζοντας όλα τα συστήματα υπολογιστών στον κόσμο .

In 1999, everybody believed that the millennium **bug** would create a global disaster by closing down computer systems across the world.

PSTS(bug, virus)

Step 3: Enumerate Sentences

x = *bug* y = *virus*

$f \in F^{xy}$	$\downarrow PMI(y, f)$
ιός (el) [virus]	11.4
virus (fr) [virus]	10.0
ميكروب (ar) [microbial]	6.5
虫 (zh) [worm]	3.4
: (fr) [<punctuation>]	-0.7

`το 1999 , όλοι πίστευαν ότι ο` **ιός** `της χιλιετίας θα προκαλούσε παγκόσμια καταστροφή επηρεάζοντας όλα τα συστήματα υπολογιστών στον κόσμο .`

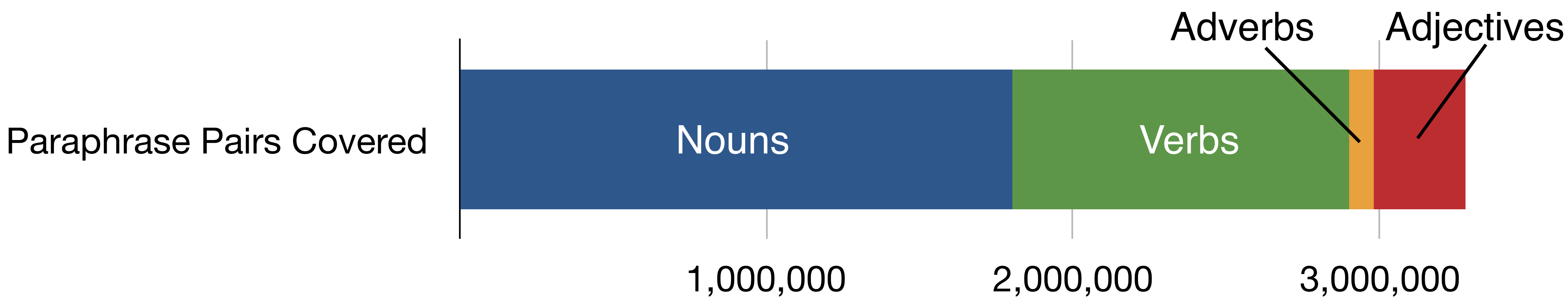
`In 1999, everybody believed that the millennium` **bug** `would create a global disaster by closing down computer systems across the world.`

PSTS(bug, virus)

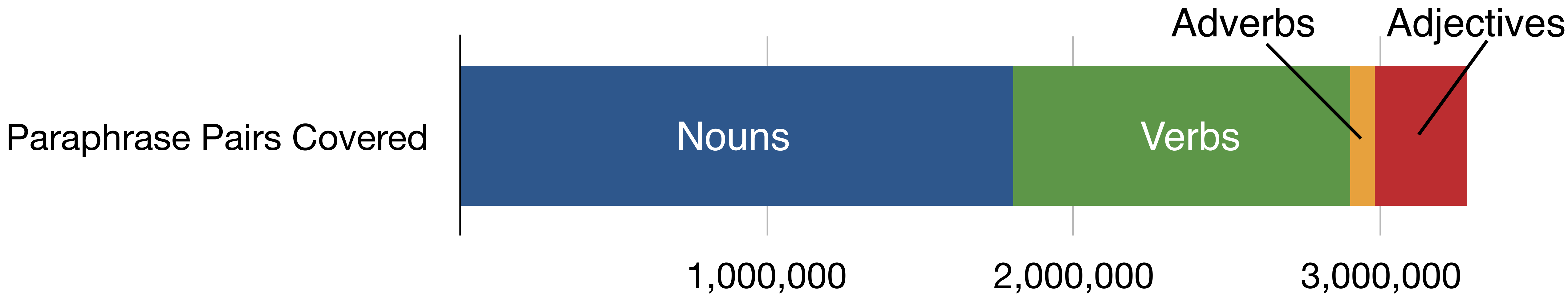
`On dirait que vous avez attrapé le` **virus** `.`

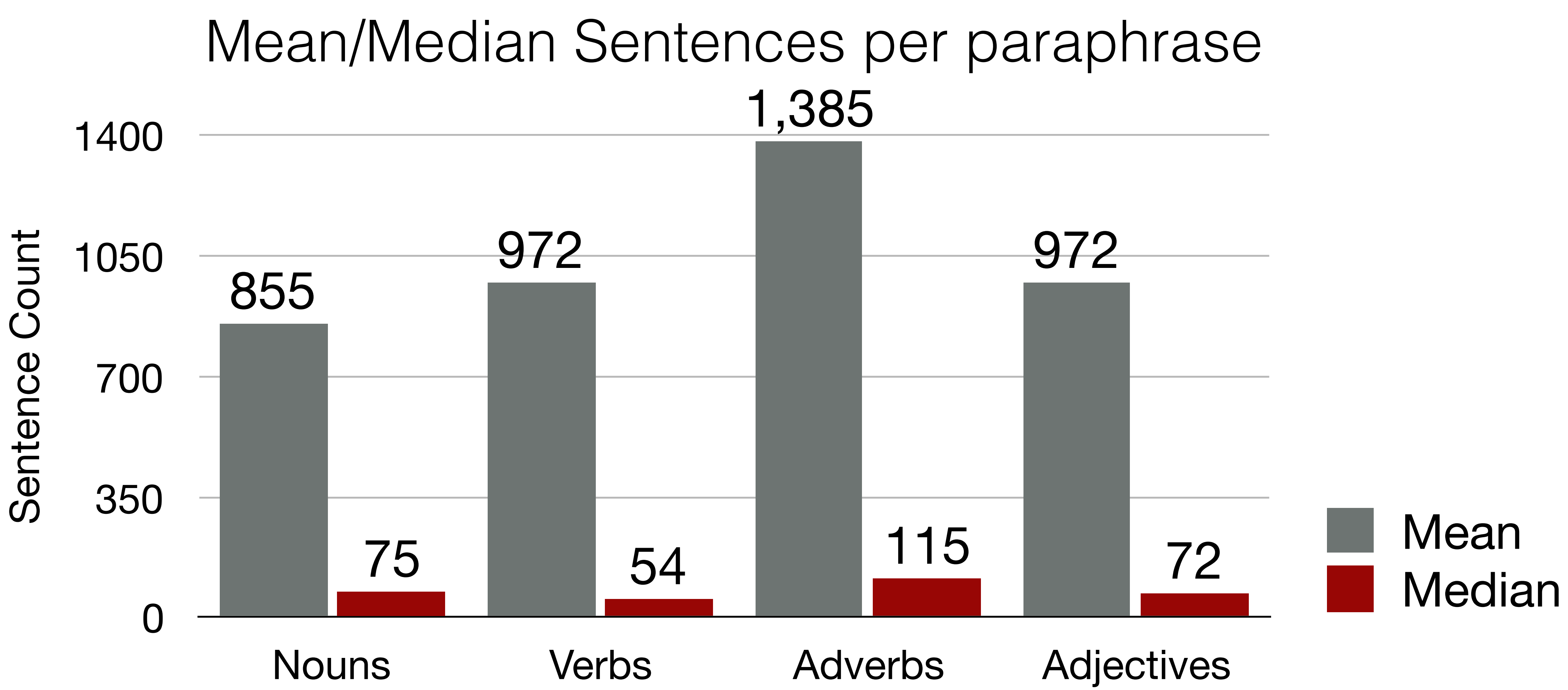
`It looks like you caught the` **bug** `.`

This method is used to extract up to 10k sentences for each of 3.3 million paraphrase pairs



This method is used to extract up to 10k sentences for each of 3.3 million paraphrase pairs





Human evaluation indicates PSTS sentences are of mixed quality…

Test Sentence	search the knowledge bases available to see if there are any documents out there describing the condition or **error** message that the system is getting .
Paraphrase	**bug**

Sometimes **error** means roughly the same thing as **bug**. Is that true in this sentence?

YES NO UNCLEAR NEVER

Human evaluation indicates PSTS sentences are of mixed quality…

Test Sentence	search the knowledge bases available to see if there are any documents out there describing the condition or **error** message that the system is getting .
Paraphrase	**bug**

Sometimes **error** means roughly the same thing as **bug**. Is that true in this sentence?

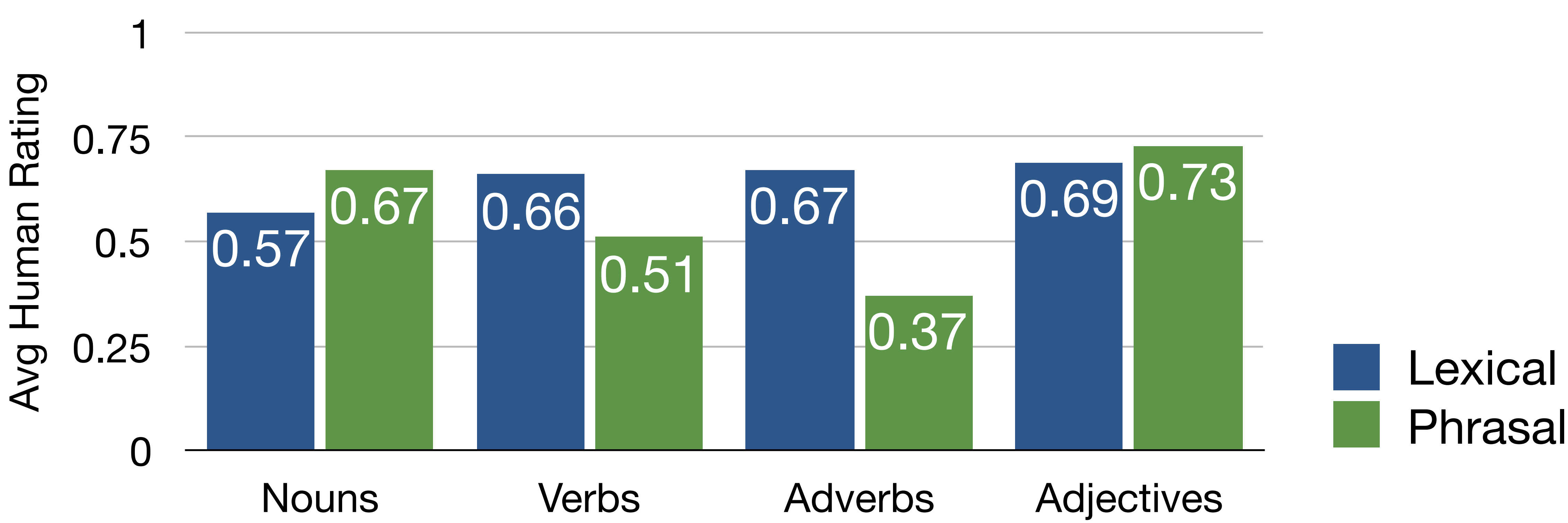
1 YES NO ~~UNCLEAR~~ NEVER

Human evaluation indicates PSTS sentences are of mixed quality…

Test Sentence	search the knowledge bases available to see if there are any documents out there describing the condition or **error** message that the system is getting .
Paraphrase	**bug**

Sometimes **error** means roughly the same thing as **bug**. Is that true in this sentence?

1 YES 0 NO ~~UNCLEAR~~ 0 NEVER

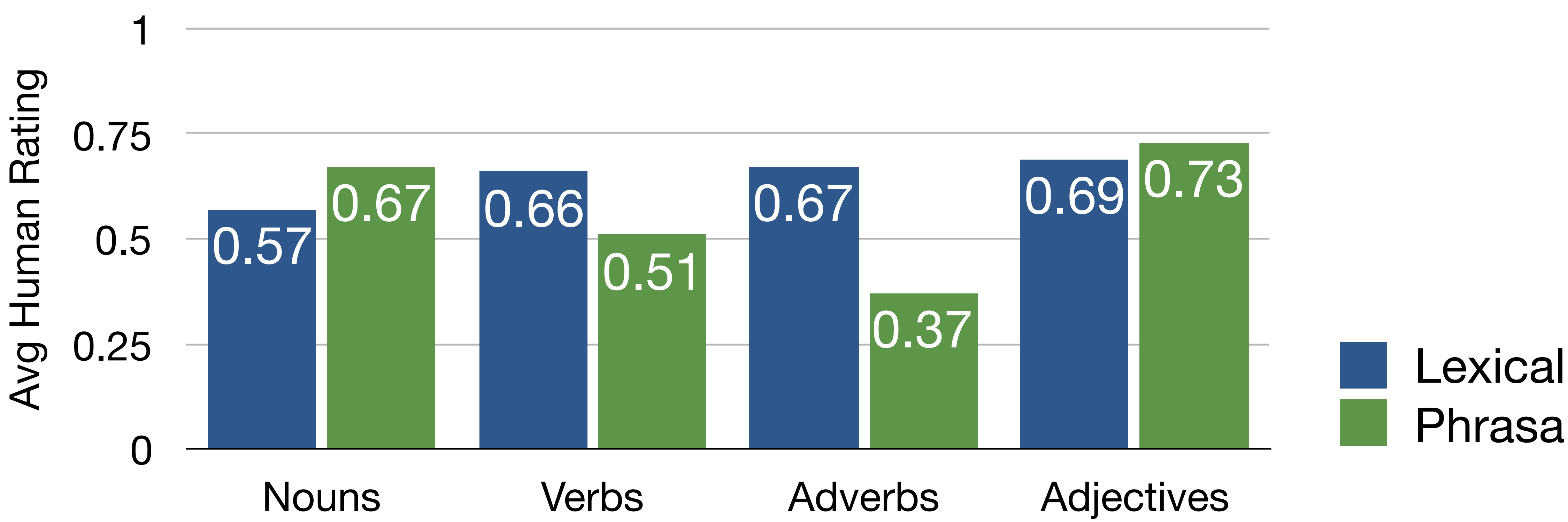


Human evaluation indicates PSTS sentences are of mixed quality…we need a way to rank sentences

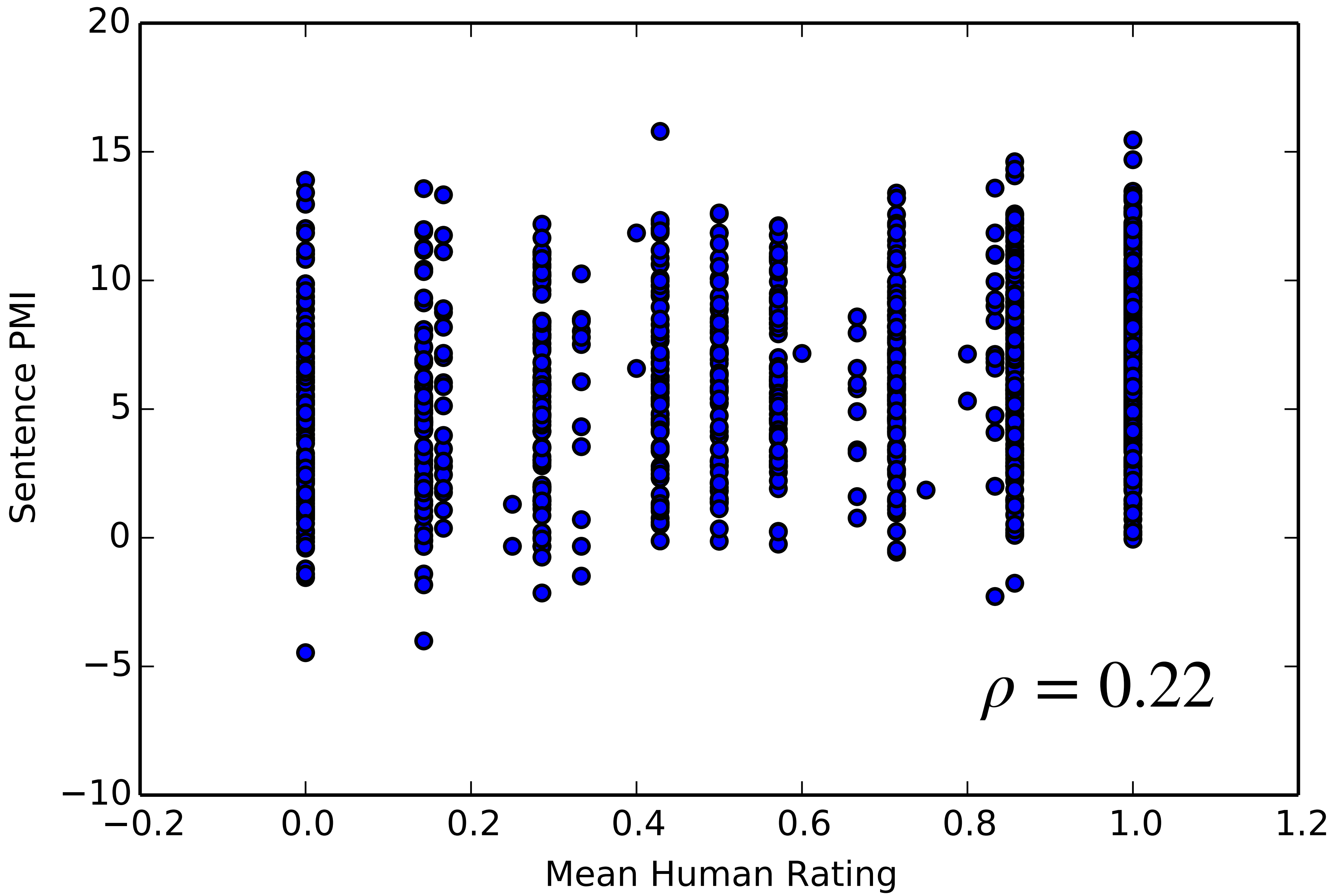
Test Sentence	search the knowledge bases available to see if there are any documents out there describing the condition or **error** message that the system is getting .
Paraphrase	**bug**

Sometimes **error** means roughly the same thing as **bug**. Is that true in this sentence?

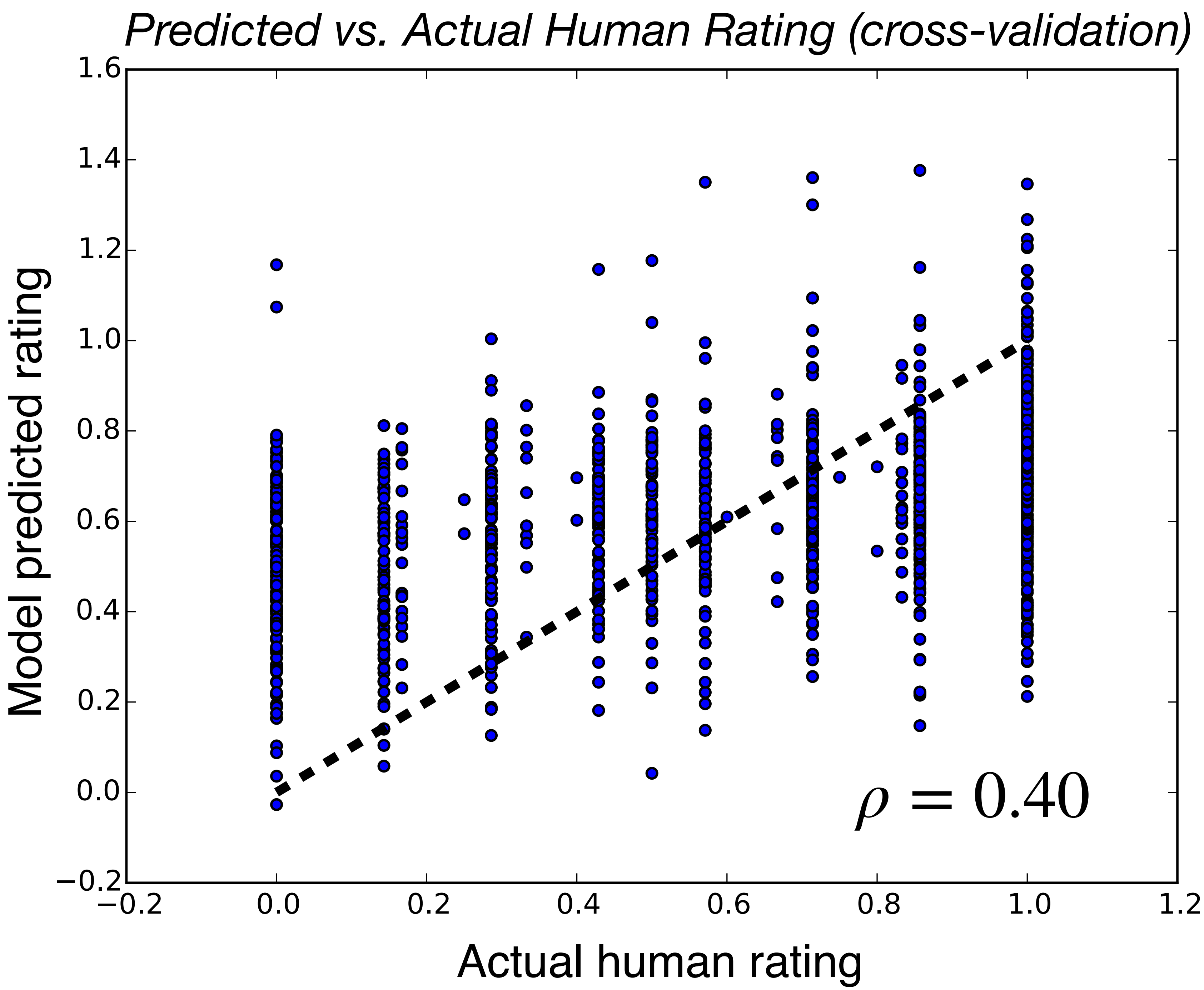
1 YES   NO   ~~UNCLEAR~~   NEVER



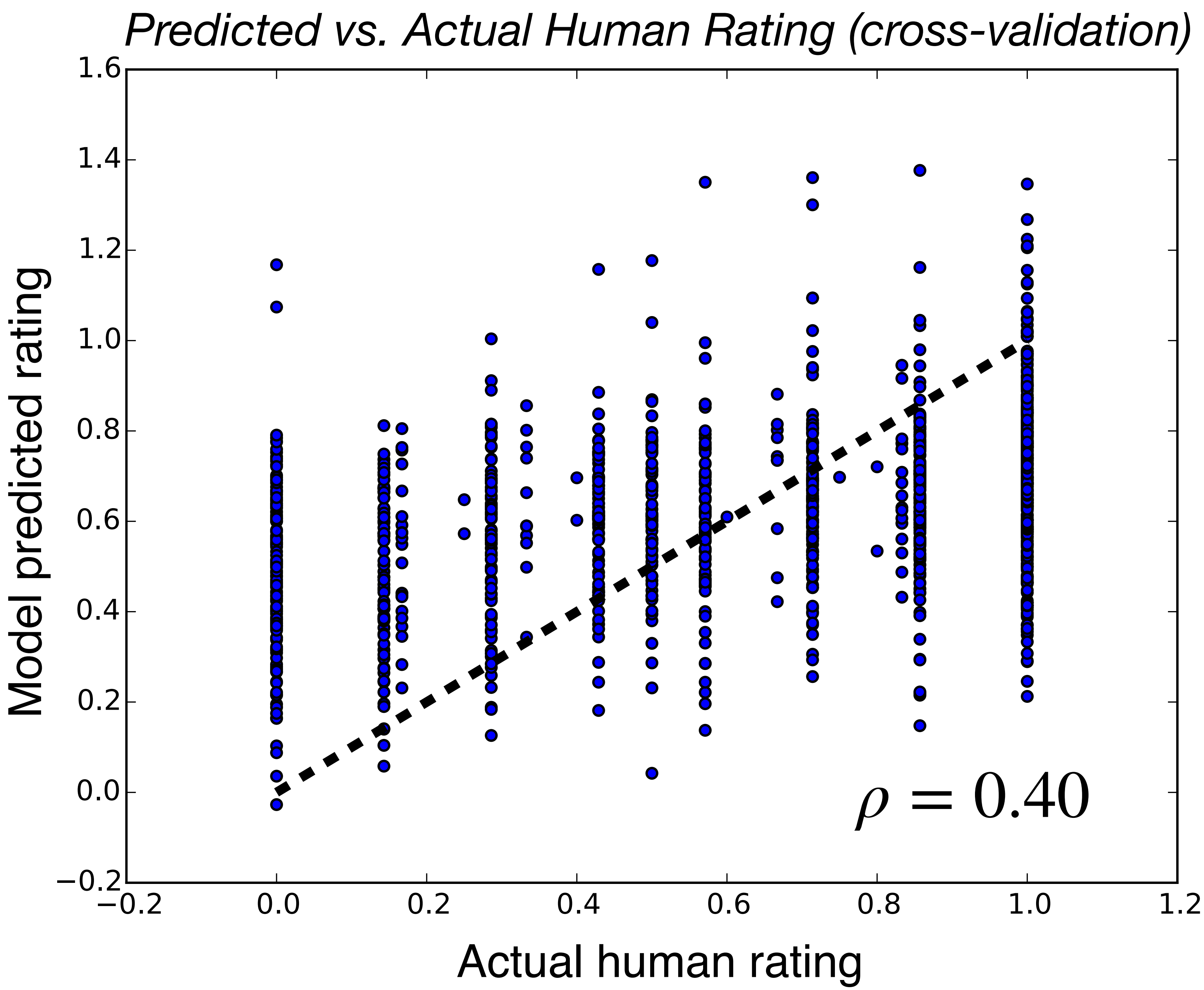
PMI is only loosely correlated with human judgments of sentence quality…



…so we train a regression model to better correlate with human judgments, which can be used to rank sentences



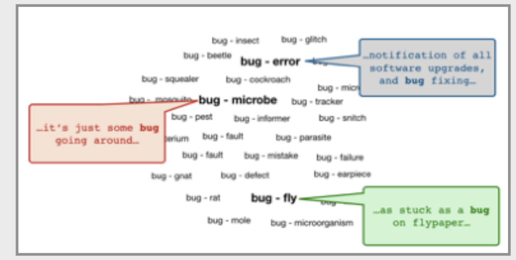
…so we train a regression model to better correlate with human judgments, which can be used to rank sentences



- Regression model predicts human rating based on input sentence and paraphrase
- Feature types
  - PPDB features
  - contextual features
  - syntactic features
  - PMI
- Training set: 1280 instances

Meaning-specific Examples of Word Use

In submission



- Claims:
  - The pivot method can be applied to generate a paraphrase-sense-tagged corpus at scale
  - The resulting resource is useful for training sense-aware models for downstream tasks

PSTS demonstrated use in three tasks

PSTS demonstrated use in three tasks

- Training word sense embeddings

WT-BERT vector	Nearest WT-BERT neighbors
v_{pest}	v_{pests} $v_{the\ pest}$ $v_{pest-control}$ $v_{pesticides}$ $v_{pesticide}$
PP-BERT vector	Nearest PP-BERT neighbors
$v_{pest \to bug}$	$v_{pest \to lice}$ $v_{pest \to cockroach}$ $v_{pest \to infection}$ $v_{pest \to larvae}$ $v_{pest \to parasite}$

PSTS demonstrated use in three tasks

- Training word sense embeddings

WT-BERT vector	Nearest WT-BERT neighbors
v_{pest}	v_{pests} $v_{the\ pest}$ $v_{pest-control}$ $v_{pesticides}$ $v_{pesticide}$
PP-BERT vector	Nearest PP-BERT neighbors
$v_{pest\rightarrow bug}$	$v_{pest\rightarrow lice}$ $v_{pest\rightarrow cockroach}$ $v_{pest\rightarrow infection}$ $v_{pest\rightarrow larvae}$ $v_{pest\rightarrow parasite}$

- Word sense induction

- Task:

sense 1

More than 1500 publishing **houses** from 38 countries and regions participated.

sense 2

The economic environment of employees buying **houses** will be eased even more.

sense 2

Members of the delegation decided to go to **houses** of farmers for a look.

PSTS demonstrated use in three tasks

- Training word sense embeddings

WT-BERT vector	Nearest WT-BERT neighbors
v_{pest}	v_{pests} $v_{the\ pest}$ $v_{pest-control}$ $v_{pesticides}$ $v_{pesticide}$
PP-BERT vector	Nearest PP-BERT neighbors
$v_{pest \to bug}$	$v_{pest \to lice}$ $v_{pest \to cockroach}$ $v_{pest \to infection}$ $v_{pest \to larvae}$ $v_{pest \to parasite}$

- Word sense induction

- Task:

sense 1
```
More than 1500 publishing houses from 38 countries and
regions participated.
```

sense 2
```
The economic environment of employees buying houses
will be eased even more.
```

sense 2
```
Members of the delegation decided to go to houses of
farmers for a look.
```

- Contextual hypernym prediction

```
(table, furniture,
 “I'm at the store buying an end table.”,
 “Furniture, furnishings, and household equipment.”,
 “YES”
)
```

PSTS can be used to develop a large dataset for training contextual hypernym prediction models

PSTS can be used to develop a large dataset for training contextual hypernym prediction models

Target Word Sentence	Related Word Sentence	Hypernym?

PSTS can be used to develop a large dataset for training contextual hypernym prediction models

Target Word Sentence	Related Word Sentence	Hypernym?
The bottom **chessboard** is the realm of cross-border transactions that occur outside of government control.	With such an unequal position on the **board**, any efforts to seek a draw are pathetic.	YES

PSTS can be used to develop a large dataset for training contextual hypernym prediction models

Target Word Sentence	Related Word Sentence	Hypernym?
The bottom **chessboard** is the realm of cross-border transactions that occur outside of government control.	With such an unequal position on the **board**, any efforts to seek a draw are pathetic.	YES
The fluting or corrugated **fiberboard** shall be firmly glued to the facings.	Industrial plants produce paper and **board** with a capacity exceeding 20 tons per day.	YES

PSTS can be used to develop a large dataset for training contextual hypernym prediction models

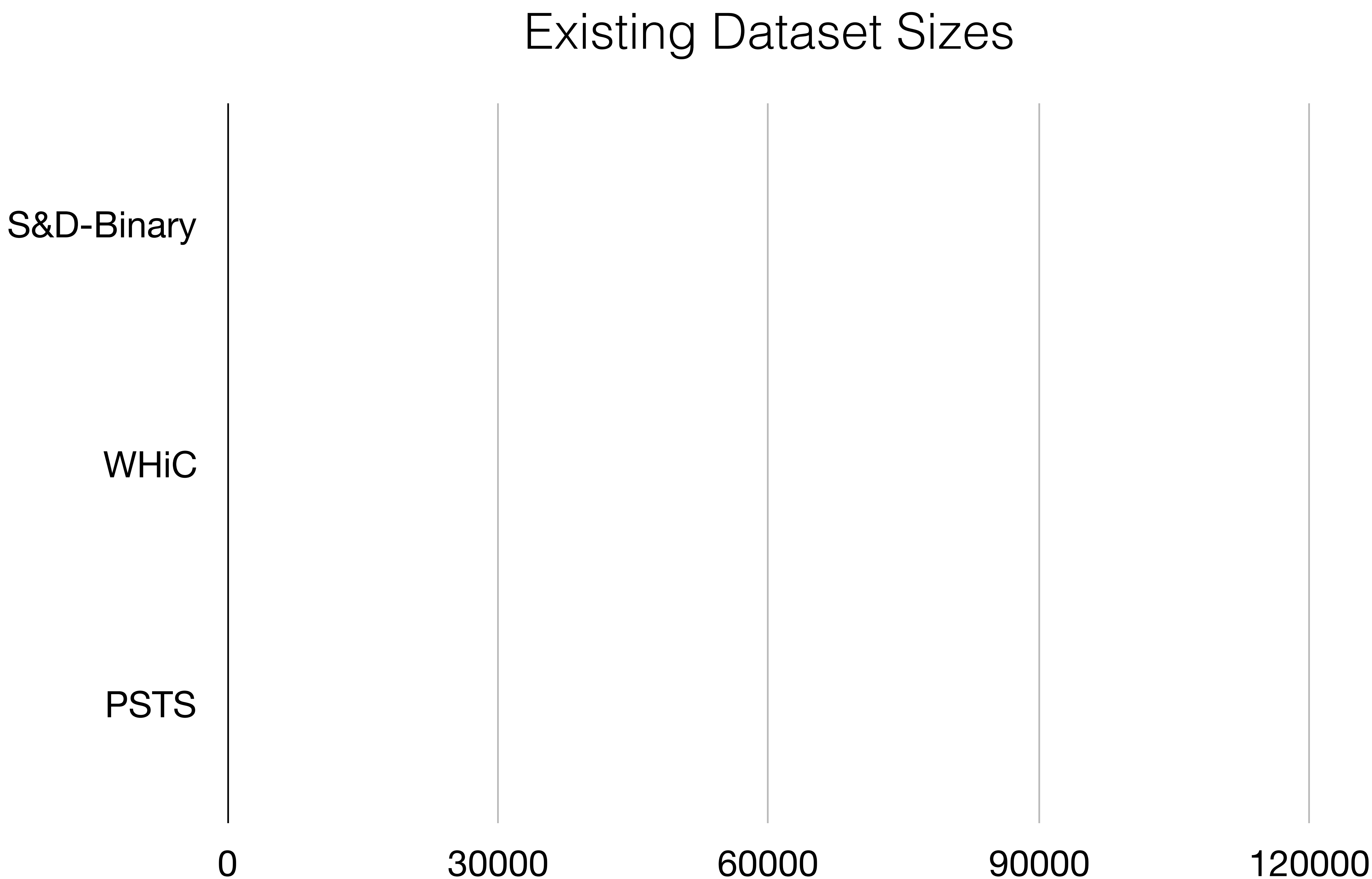
Target **Word Sentence**	**Related** **Word Sentence**	**Hypernym?**
The bottom **chessboard** is the realm of cross-border transactions that occur outside of government control.	With such an unequal position on the **board**, any efforts to seek a draw are pathetic.	YES
The fluting or corrugated **fiberboard** shall be firmly glued to the facings.	Industrial plants produce paper and **board** with a capacity exceeding 20 tons per day.	YES
The bottom **chessboard** is the realm of cross-border transactions that occur outside of government control.	These people are already on **board** fishing vessels.	NO

PSTS can be used to develop a large dataset for training contextual hypernym prediction models

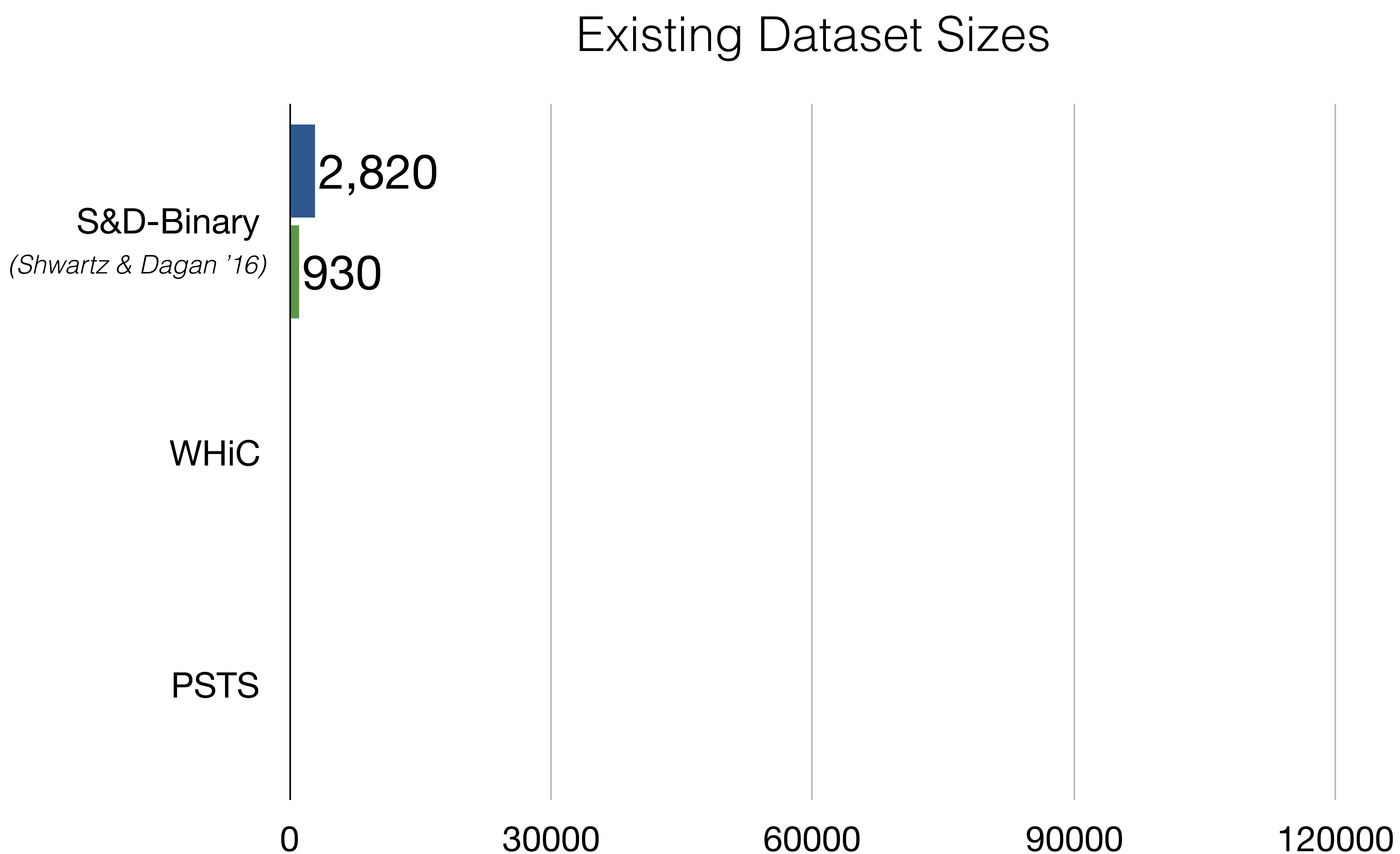
Target Word Sentence	**Related** Word Sentence	Hypernym?
The bottom **chessboard** is the realm of cross-border transactions that occur outside of government control.	With such an unequal position on the **board**, any efforts to seek a draw are pathetic.	YES
The fluting or corrugated **fiberboard** shall be firmly glued to the facings.	Industrial plants produce paper and **board** with a capacity exceeding 20 tons per day.	YES
The bottom **chessboard** is the realm of cross-border transactions that occur outside of government control.	These people are already on **board** fishing vessels.	NO
t, c_t	w, c_w	y

PSTS can be used to develop a large dataset for training contextual hypernym prediction models

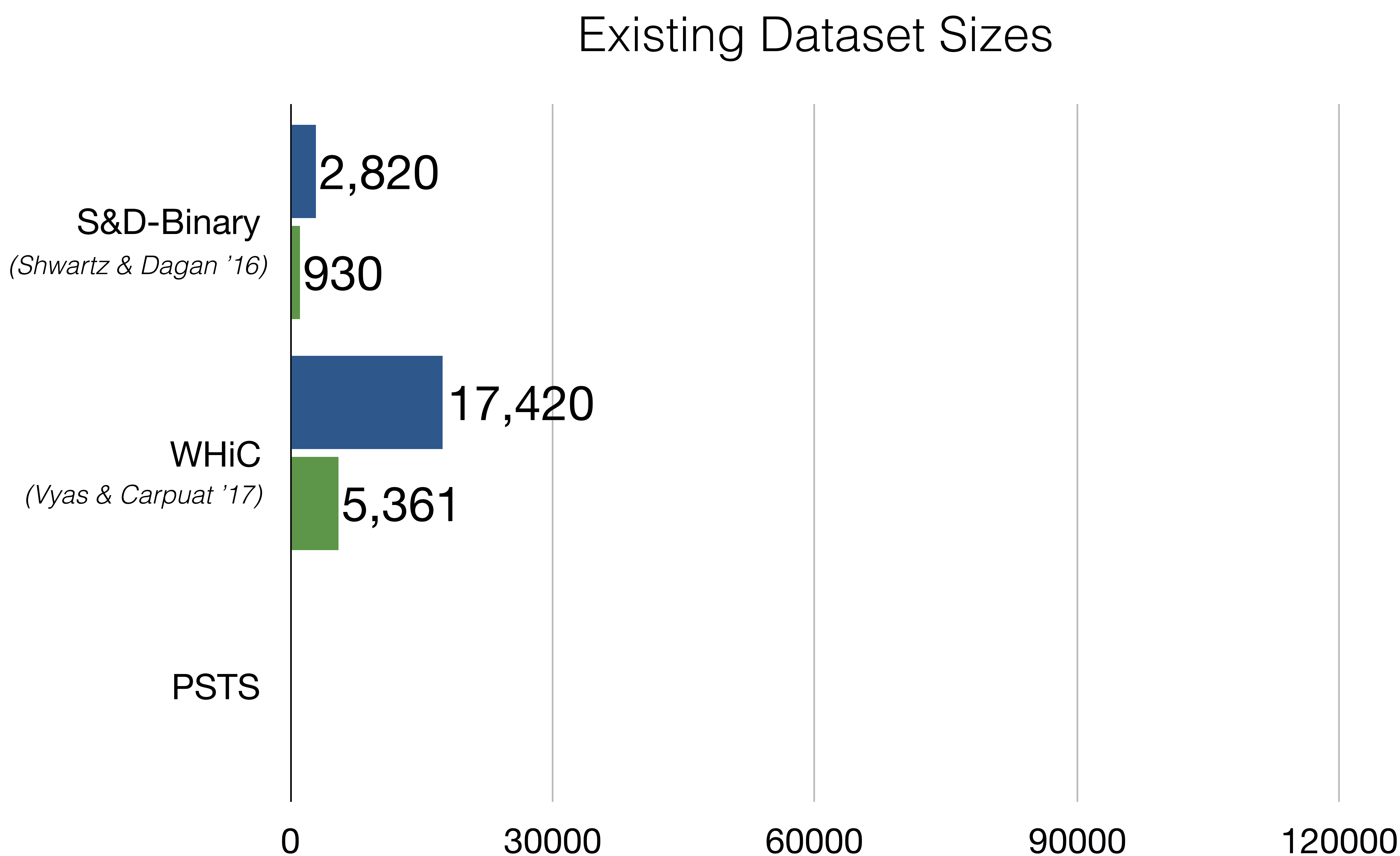
PSTS can be used to develop a large dataset for training contextual hypernym prediction models



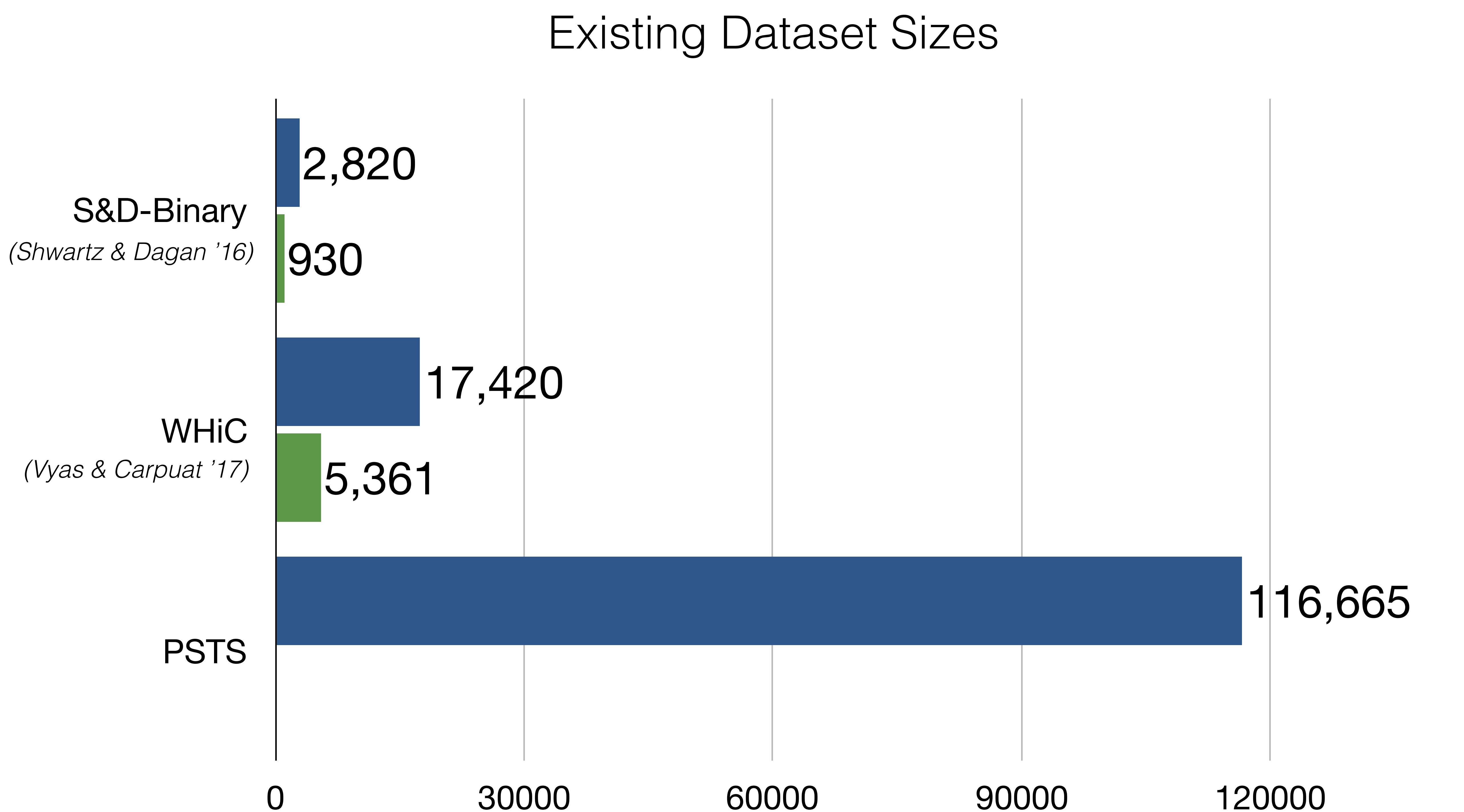
PSTS can be used to develop a large dataset for training contextual hypernym prediction models



PSTS can be used to develop a large dataset for training contextual hypernym prediction models



PSTS can be used to develop a large dataset for training contextual hypernym prediction models



PSTS can be used to develop a large dataset for training contextual hypernym prediction models

1. Find related terms in $PSTS \cap WordNet$:

```
(table, furniture)
```
hypernym

```
(table, leg)
```
meronym

PSTS can be used to develop a large dataset for training contextual hypernym prediction models

`(table, furniture)` hypernym

2 Generate related instances

$$\texttt{table} \rightarrow t$$
$$\texttt{furniture} \rightarrow w$$
$$s_i \in PSTS(table, furniture) \rightarrow c_t$$
$$s_j \in PSTS(furniture, table) \rightarrow c_w$$
$$\texttt{YES} \rightarrow y$$

PSTS can be used to develop a large dataset for training contextual hypernym prediction models

`(table, furniture)` hypernym

2 Generate related instances

$$\texttt{table} \rightarrow t$$
$$\texttt{furniture} \rightarrow w$$
$$s_i \in PSTS(table, furniture) \rightarrow c_t$$
$$s_j \in PSTS(furniture, table) \rightarrow c_w$$
$$\texttt{YES} \rightarrow y$$

```
(table, furniture,
 “I’m at the store buying an end table.”,
 “Furniture, furnishings, and household equipment.”,
 “YES”
)
```

PSTS can be used to develop a large dataset for training contextual hypernym prediction models

`(table, leg)` meronym

2 Generate related instances

$$\texttt{table} \rightarrow t$$
$$\texttt{leg} \rightarrow w$$
$$s_i \in PSTS(table, leg) \rightarrow c_t$$
$$s_j \in PSTS(leg, table) \rightarrow c_w$$
$$\texttt{NO} \rightarrow y$$

PSTS can be used to develop a large dataset for training contextual hypernym prediction models

`(table, leg)` meronym

2 Generate related instances

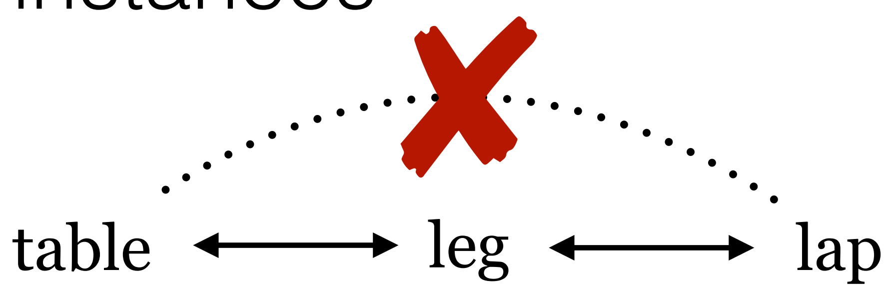
$$\texttt{table} \rightarrow t$$
$$\texttt{leg} \rightarrow w$$
$$s_i \in PSTS(table, leg) \rightarrow c_t$$
$$s_j \in PSTS(leg, table) \rightarrow c_w$$
$$\texttt{NO} \rightarrow y$$

```
(table, leg,
 "Set the plates on the table for me, please.",
 "It got a scratch in the leg during shipment.",
 "NO"
)
```

PSTS can be used to develop a large dataset for training contextual hypernym prediction models

`(table, leg)` meronym

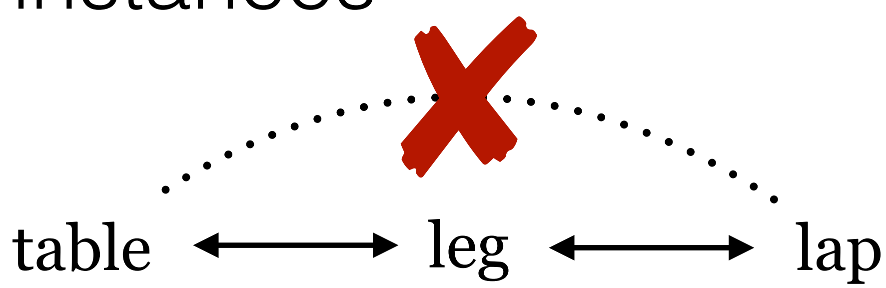
3 Generate unrelated instances



PSTS can be used to develop a large dataset for training contextual hypernym prediction models

(table, leg) meronym

3 Generate unrelated instances

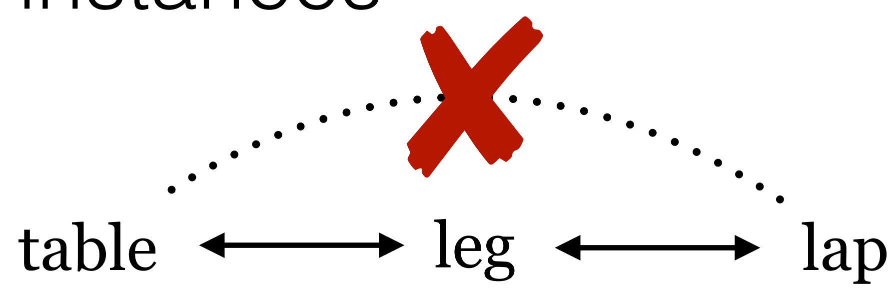


$$\texttt{table} \rightarrow t$$
$$\texttt{leg} \rightarrow w$$
$$s_i \in PSTS(table, leg) \rightarrow c_t$$
$$s_j \in PSTS(leg, lap) \rightarrow c_w$$
$$\texttt{NO} \rightarrow y$$

PSTS can be used to develop a large dataset for training contextual hypernym prediction models

`(table, leg)` meronym

3 Generate unrelated instances



$$\texttt{table} \rightarrow t$$
$$\texttt{leg} \rightarrow w$$
$$s_i \in PSTS(table, leg) \rightarrow c_t$$
$$s_j \in PSTS(leg, lap) \rightarrow c_w$$
$$\text{NO} \rightarrow y$$

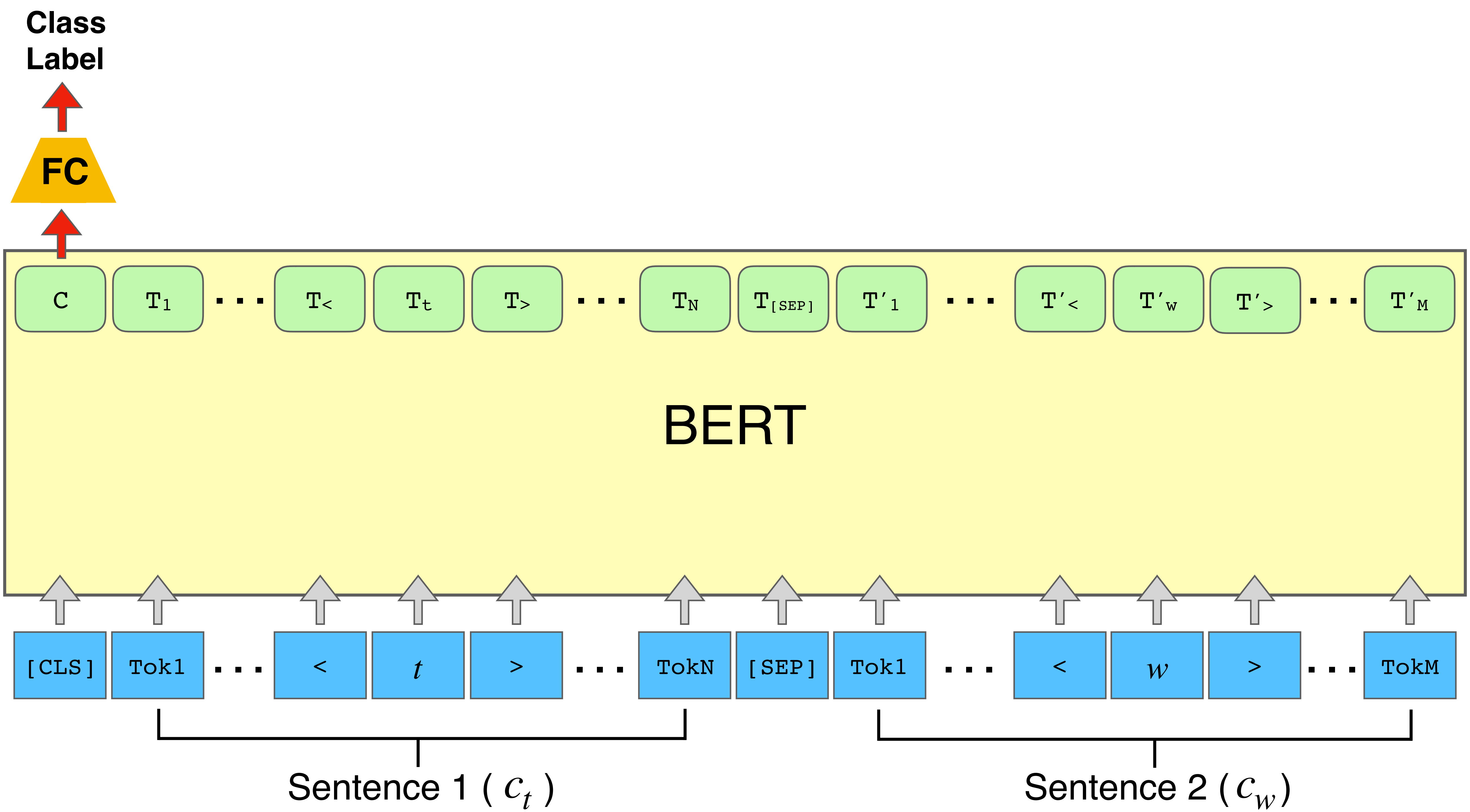
```
(table, leg,
 "Set the plates on the table for me, please.",
 "She hit a wall during the last leg of the race.",
 "NO"
)
```

Experiments

Experiments

- Evaluate performance of hypernym prediction models trained on **PSTS** vs. **S&D-binary** vs. **WHiC**
- Test on existing S&D-binary, WHiC test sets
- Model: BERT

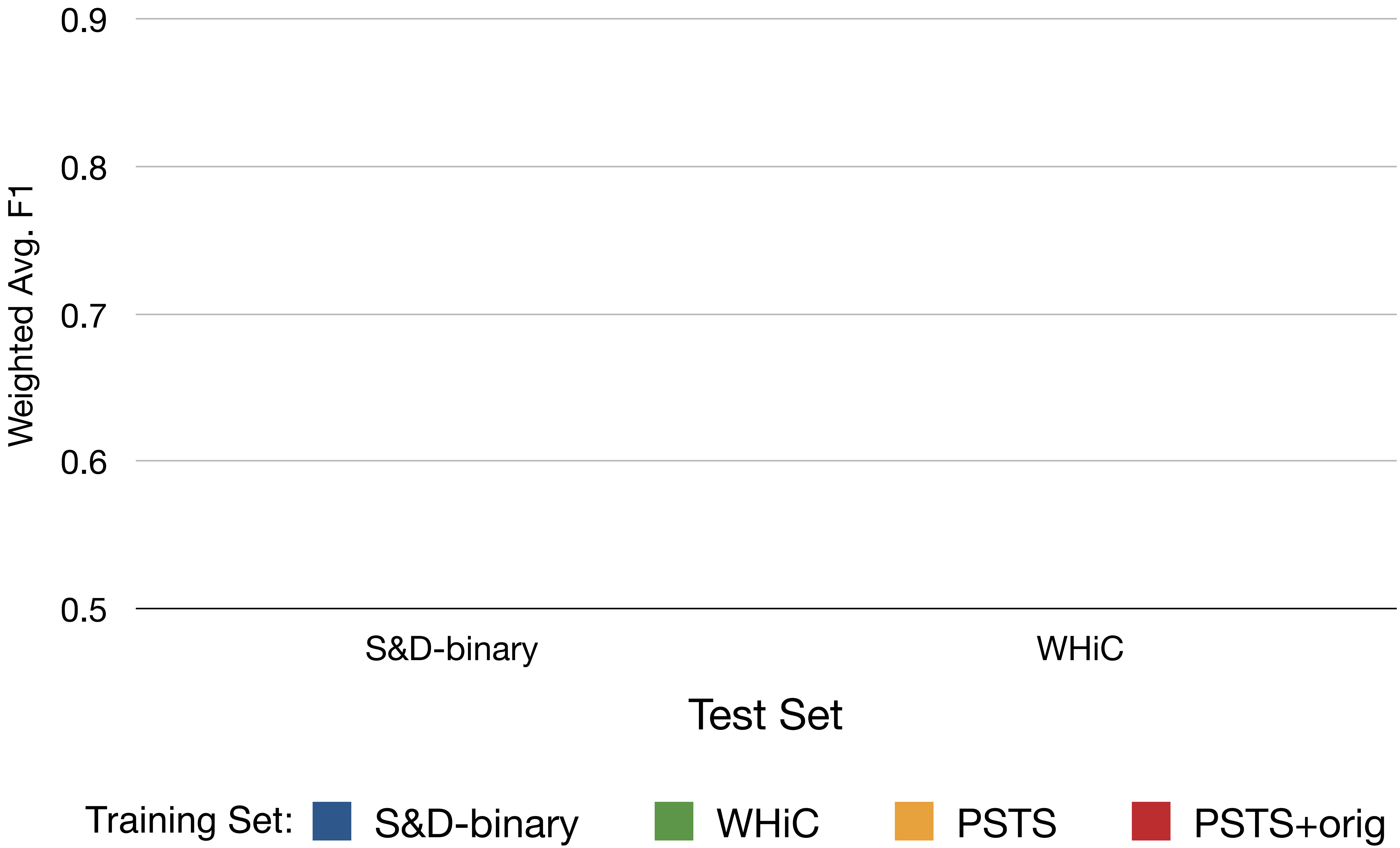
The BERT transformer encoder can be fine-tuned for the contextual hypernym prediction task



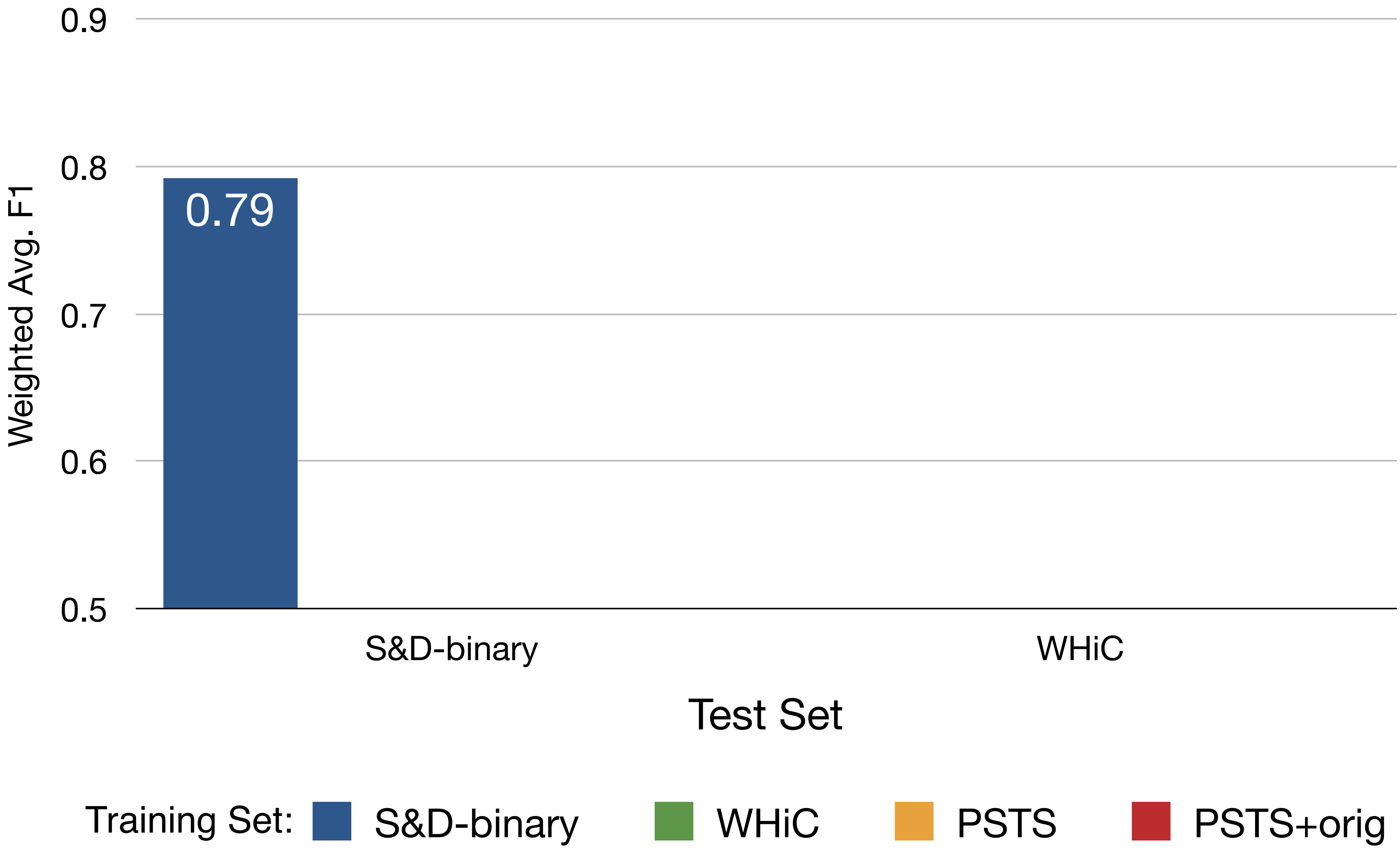
Testing models on S&D-binary: Larger PSTS training set produces better results than hand-crafted S&D-binary

Training Set:

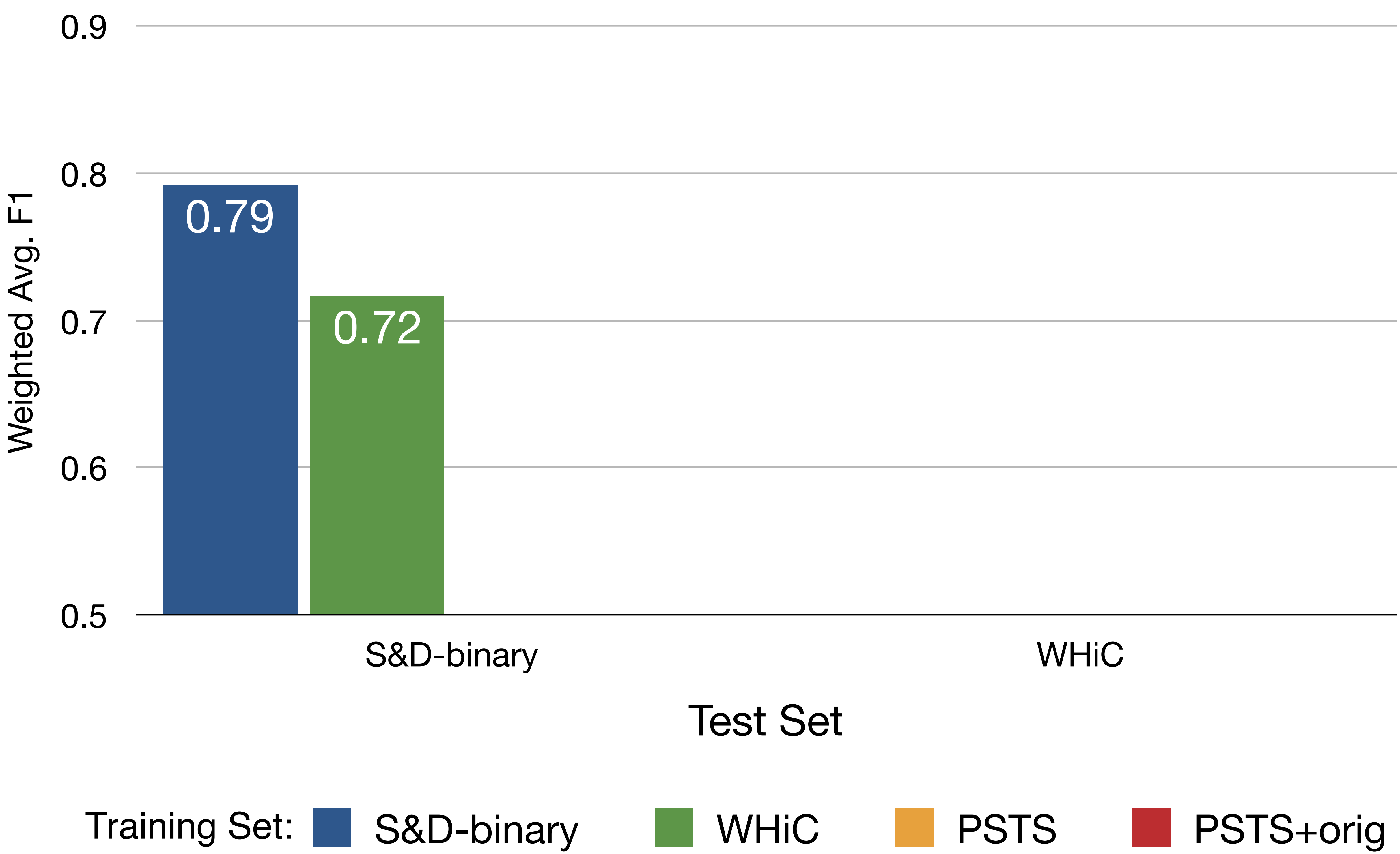
Testing models on S&D-binary: Larger PSTS training set produces better results than hand-crafted S&D-binary



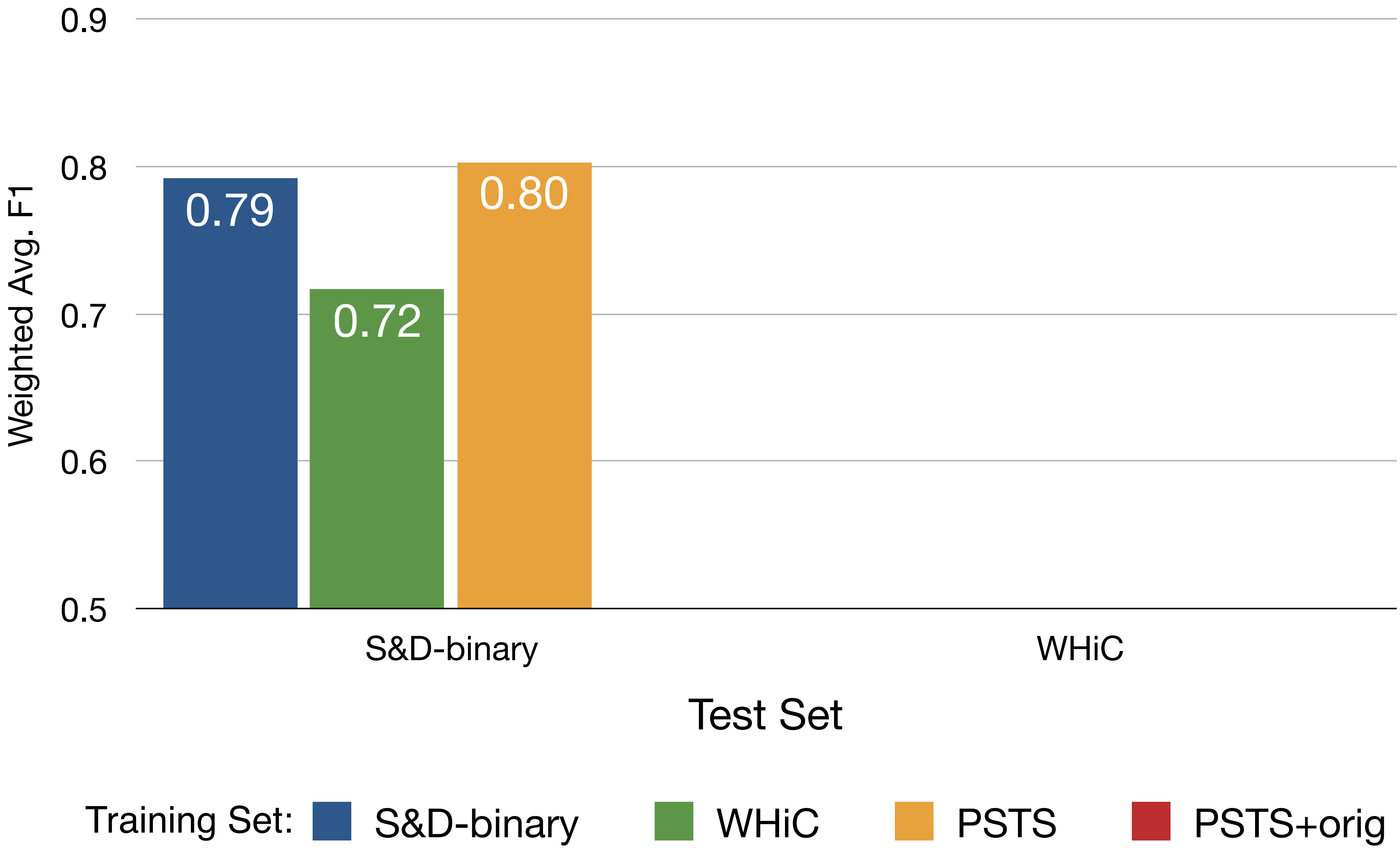
Testing models on S&D-binary: Larger PSTS training set produces better results than hand-crafted S&D-binary



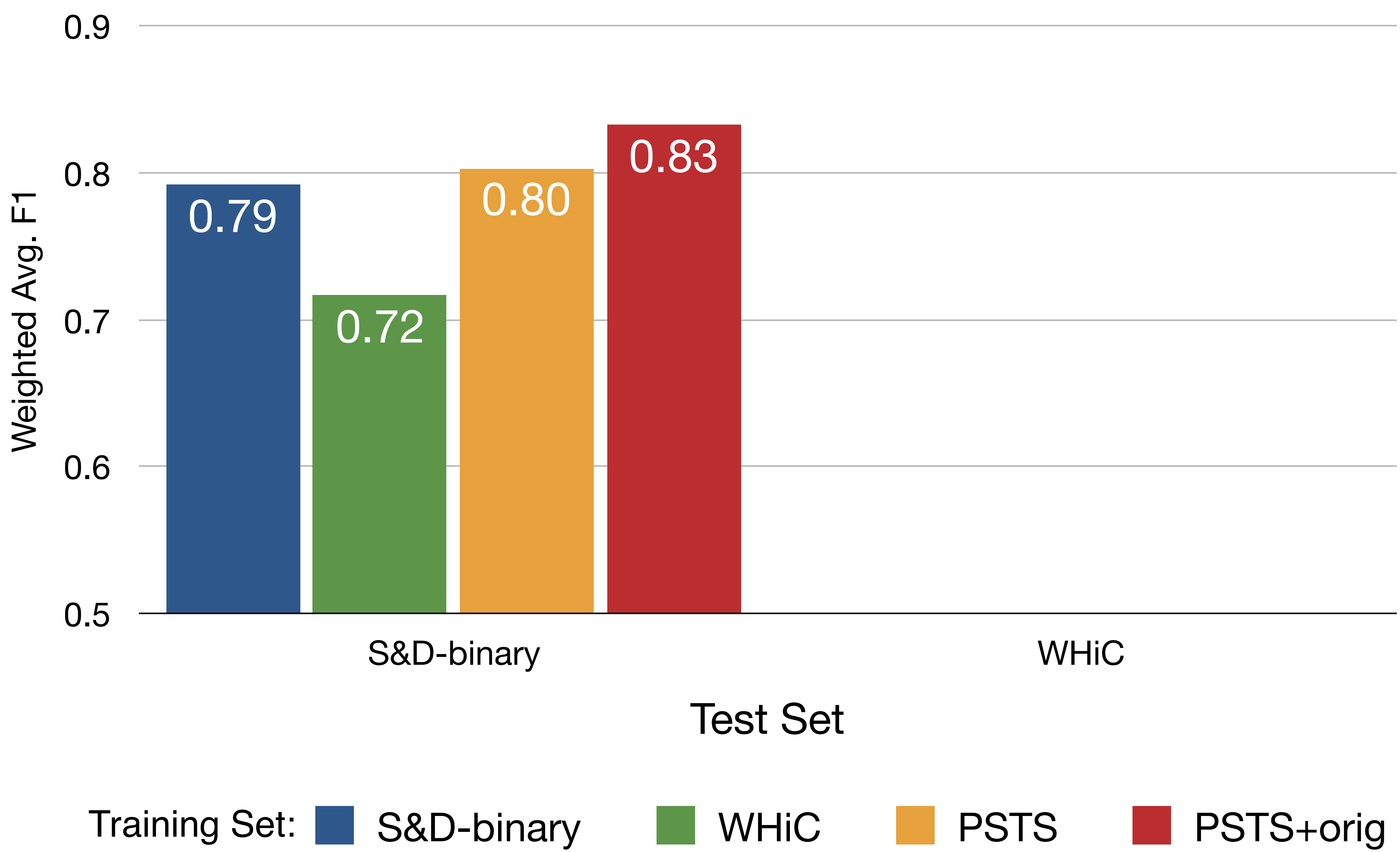
Testing models on S&D-binary: Larger PSTS training set produces better results than hand-crafted S&D-binary



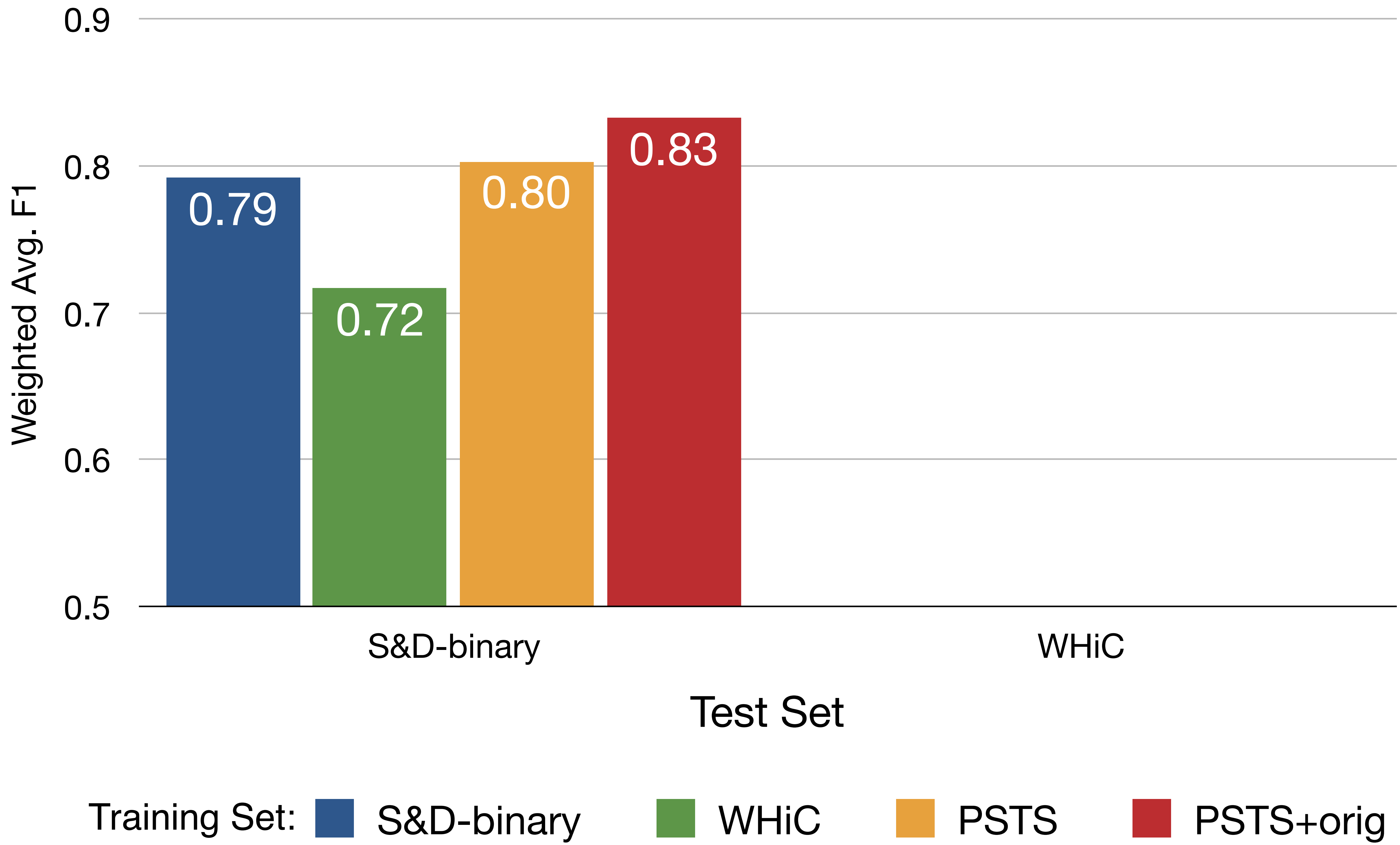
Testing models on S&D-binary: Larger PSTS training set produces better results than hand-crafted S&D-binary



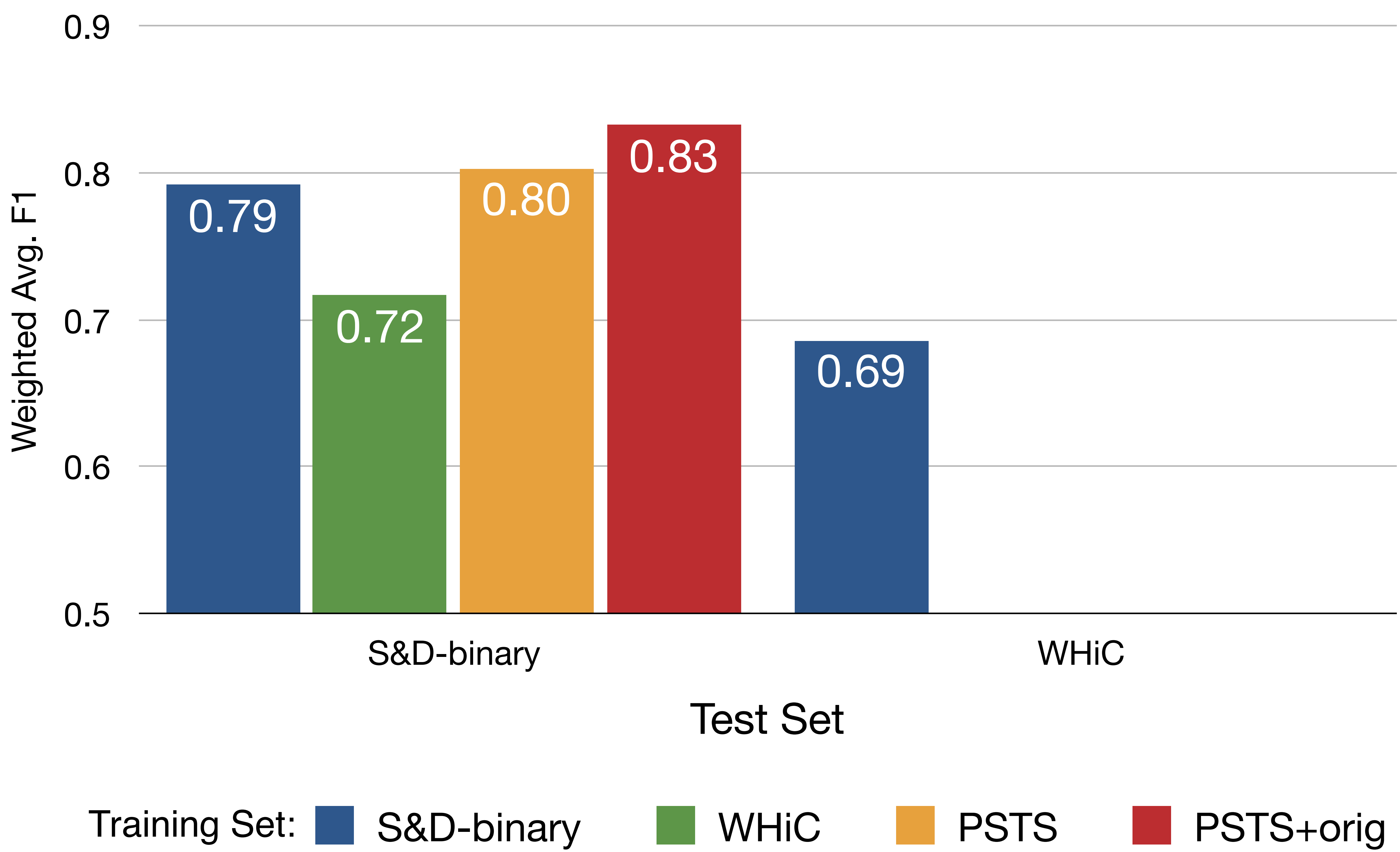
Testing models on S&D-binary: Larger PSTS training set produces better results than hand-crafted S&D-binary



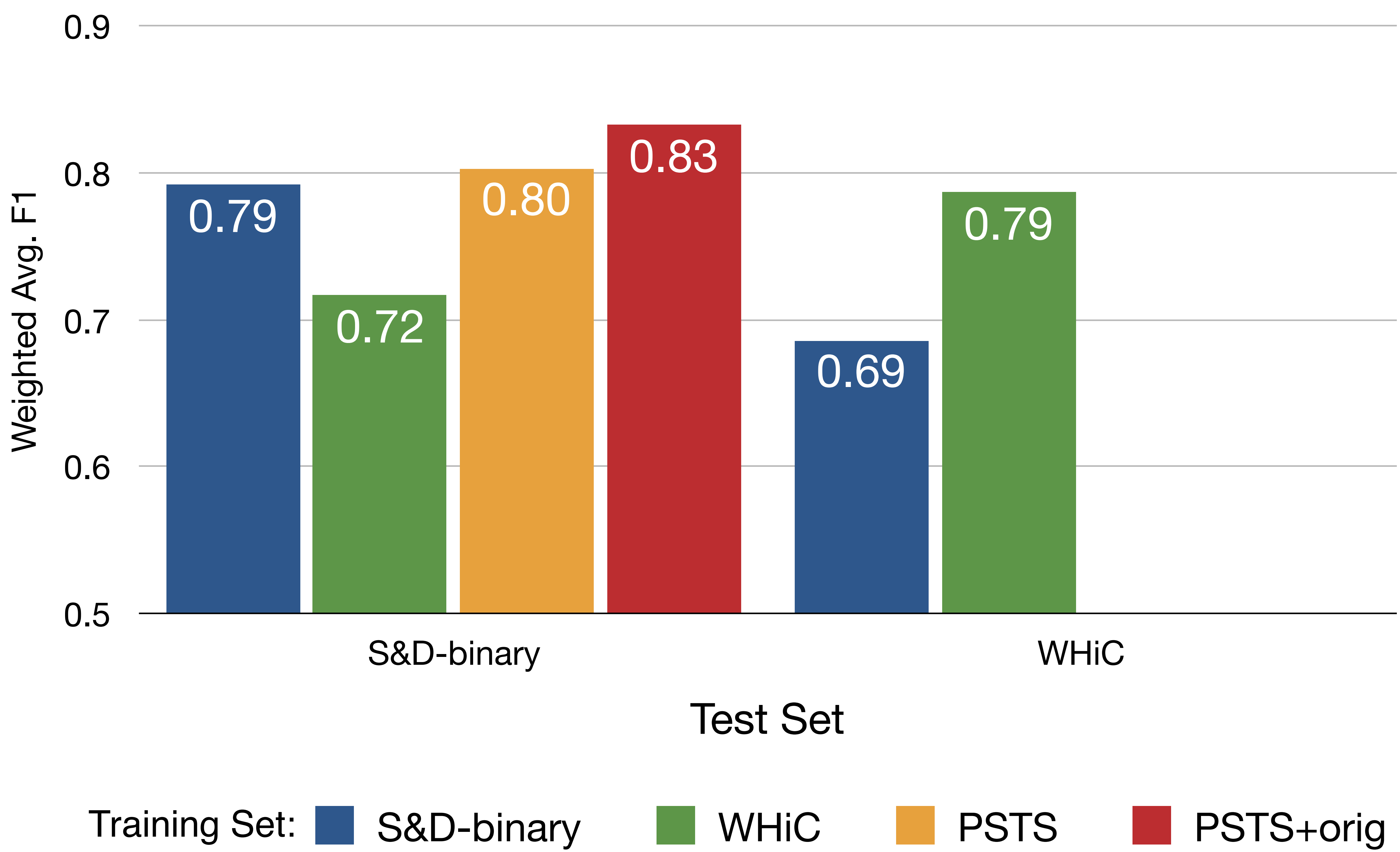
Testing models on WHiC: Larger PSTS training set does not produce better results than original training set



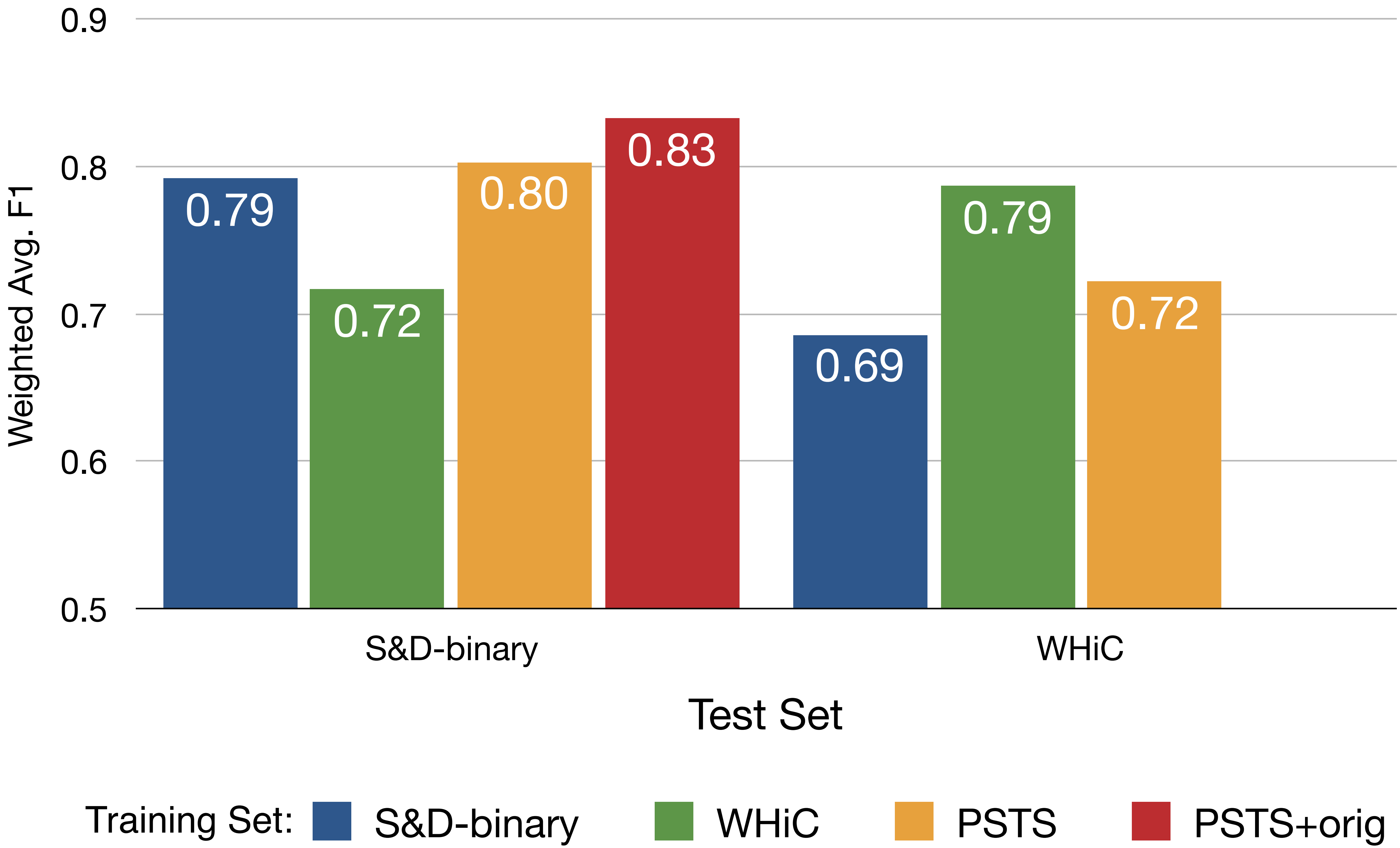
Testing models on WHiC: Larger PSTS training set does not produce better results than original training set



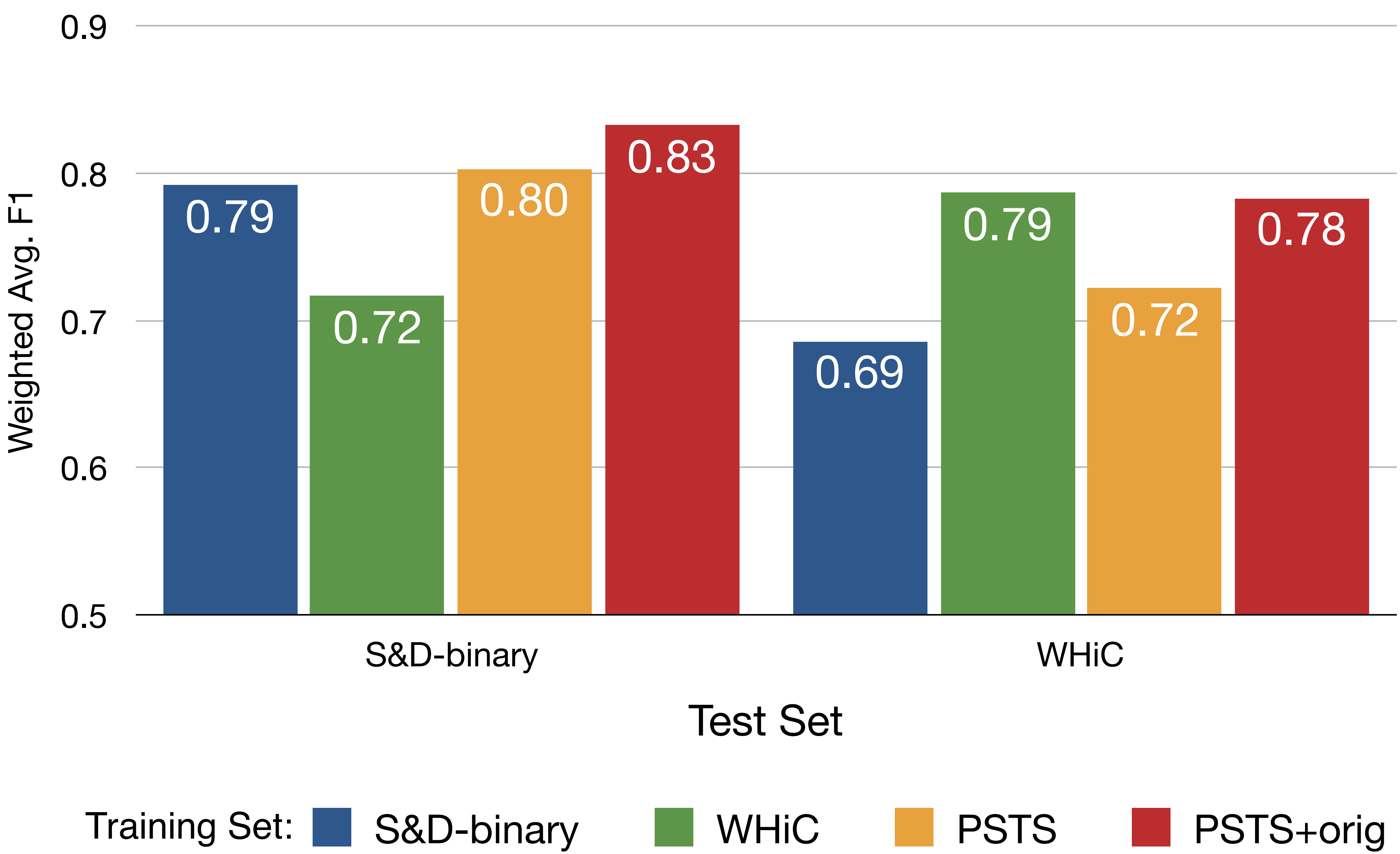
Testing models on WHiC: Larger PSTS training set does not produce better results than original training set



Testing models on WHiC: Larger PSTS training set does not produce better results than original training set

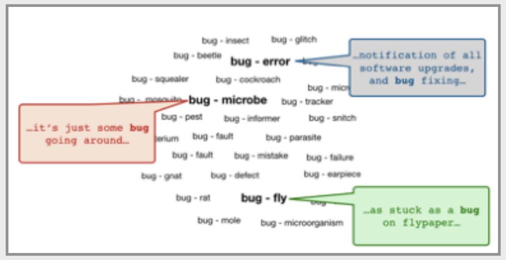


Testing models on WHiC: Larger PSTS training set does not produce better results than original training set



Meaning-specific Examples of Word Use

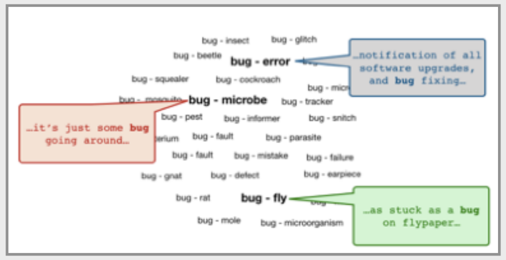
In submission



- Claims:
  - The pivot method can be applied to generate a paraphrase-sense-tagged corpus at scale
  - The resulting resource is useful for training sense-aware models for downstream tasks

Meaning-specific Examples of Word Use

In submission



- Take-aways:
  - Paraphrases-as-senses is a useful abstraction for modeling fine-grained word meaning
  - Paraphrases are a similar, but alternative, method to foreign translations for automatically generating sense-tagged corpora

Motivation

Using Paraphrases to Model Word Sense

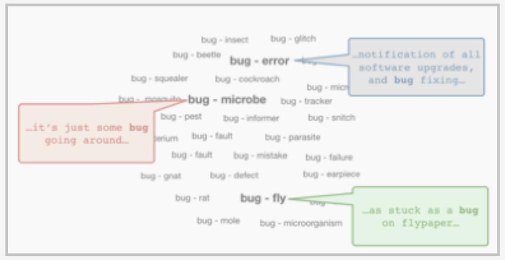
NAACL 2016; SENSE@EACL 2017

Learning Scalar Adjective Intensity

EMNLP 2018

Meaning-specific Examples of Word Use

In submission



Conclusion & Future Work

Motivation

Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

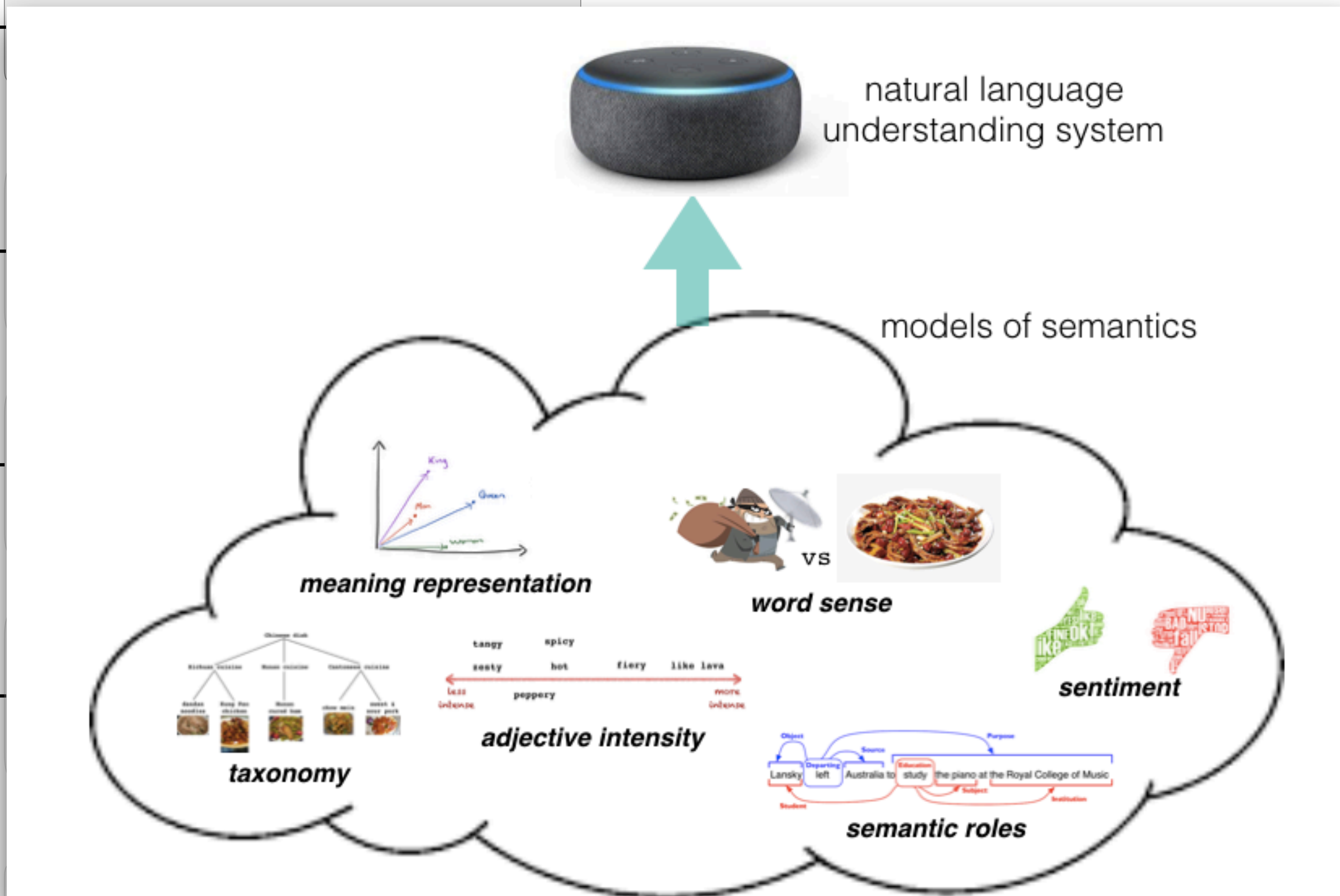
Learning Scalar Adjective Intensity

EMNLP 2018

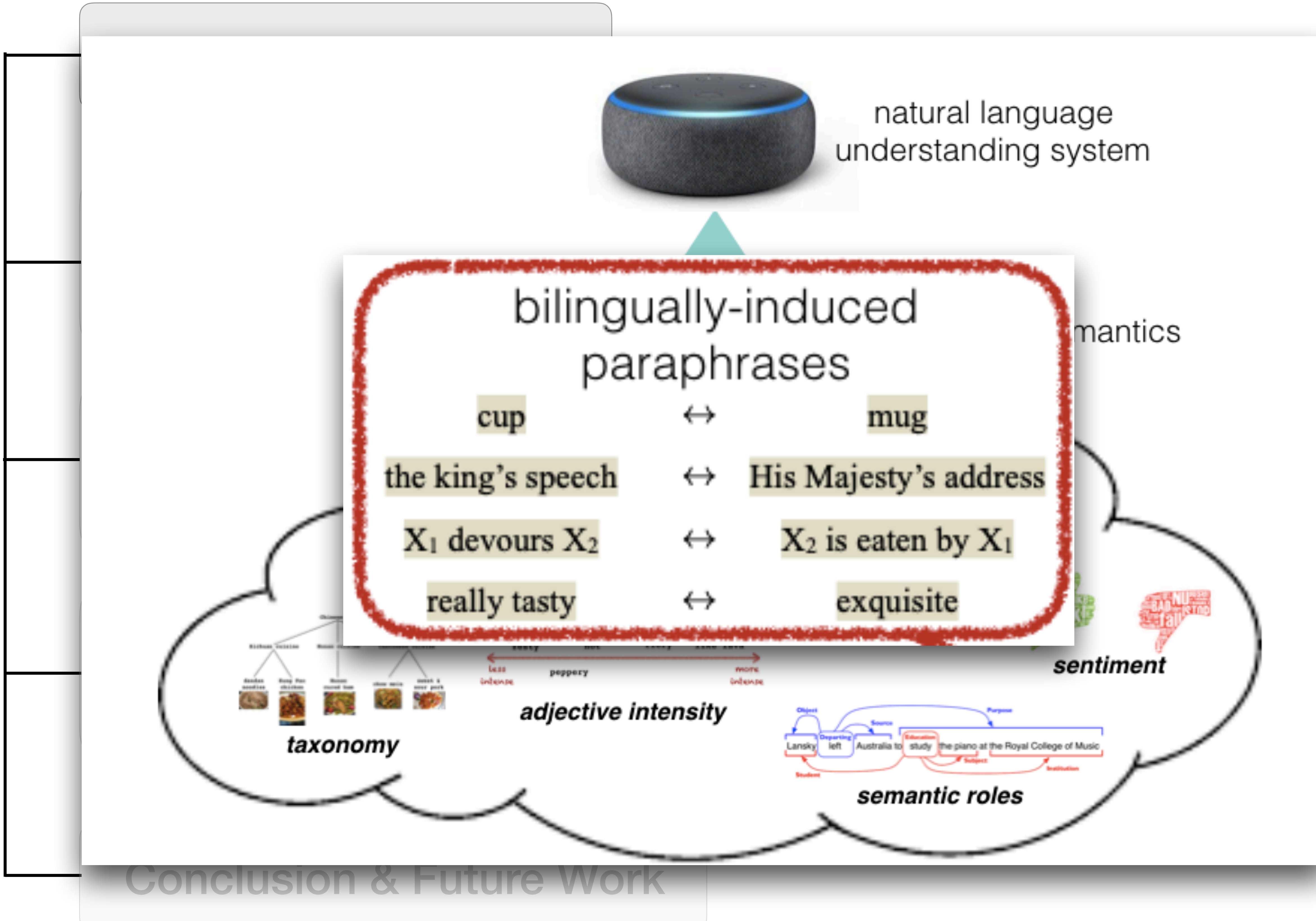
Meaning-specific Examples of Word Use

In submission

Conclusion & Future Work



Conclusion & Future Work


natural language
understanding system
mantics
bilingually-induced
paraphrases
cup ↔ mug
the king's speech ↔ His Majesty's address
X_1 devours X_2 ↔ X_2 is eaten by X_1
really tasty ↔ exquisite
taxonomy
adjective intensity
sentiment
semantic roles
Conclusion & Future Work

Motivation

Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

Learning Scalar Adjective Intensity

EMNLP 2018

Meaning-specific Examples of Word Use

In submission

Conclusion & Future Work

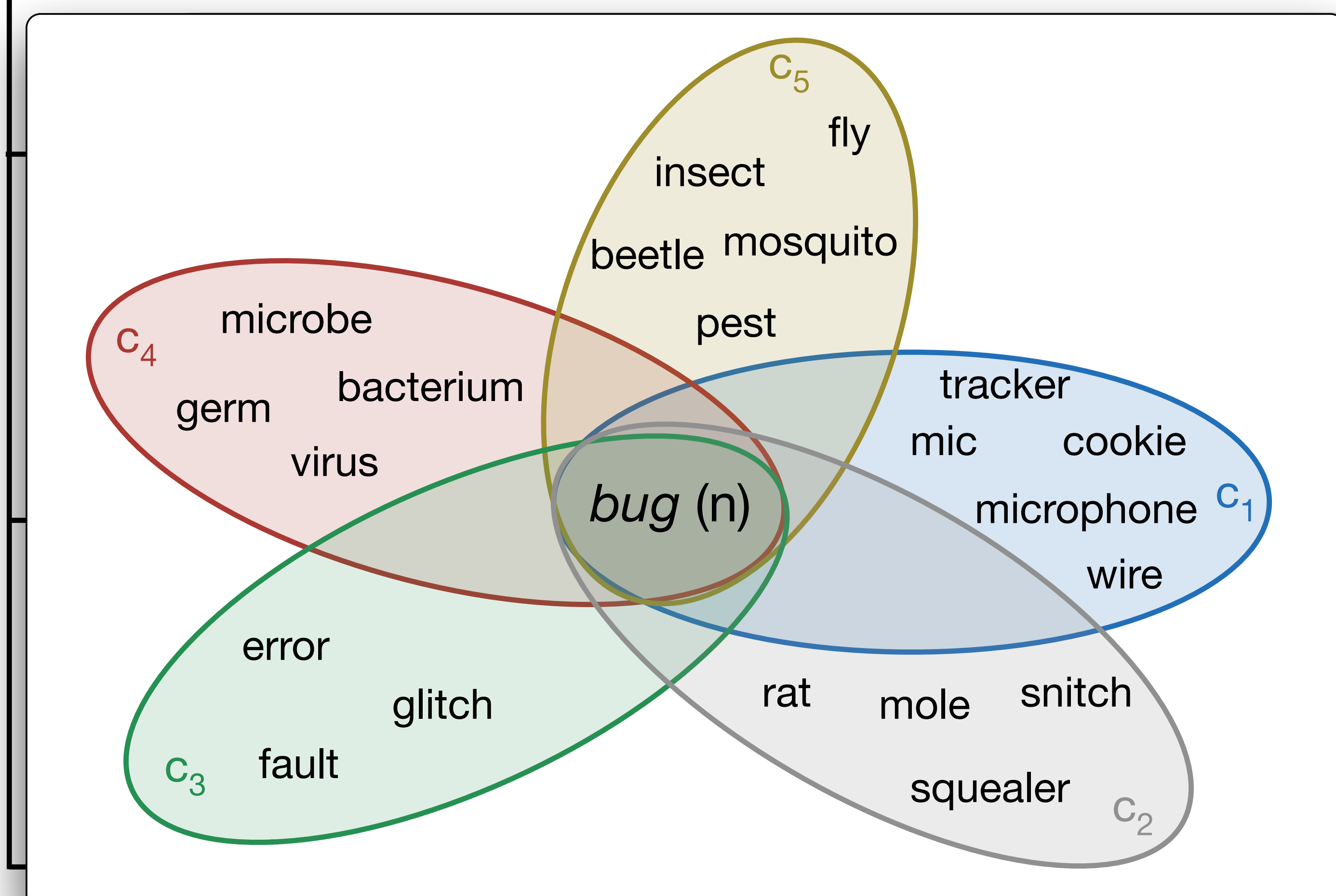
Motivation

Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

...sity

...Word Use

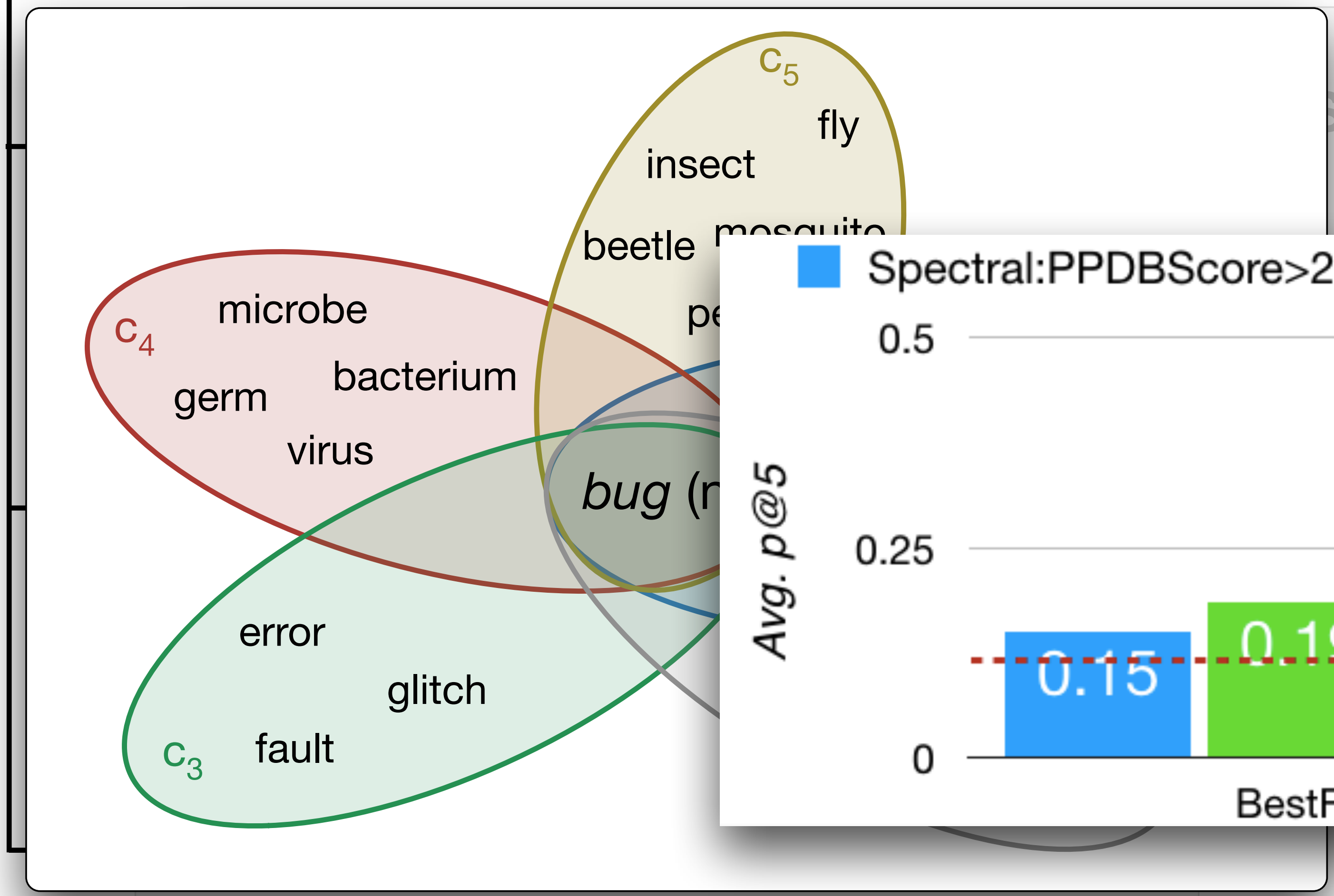


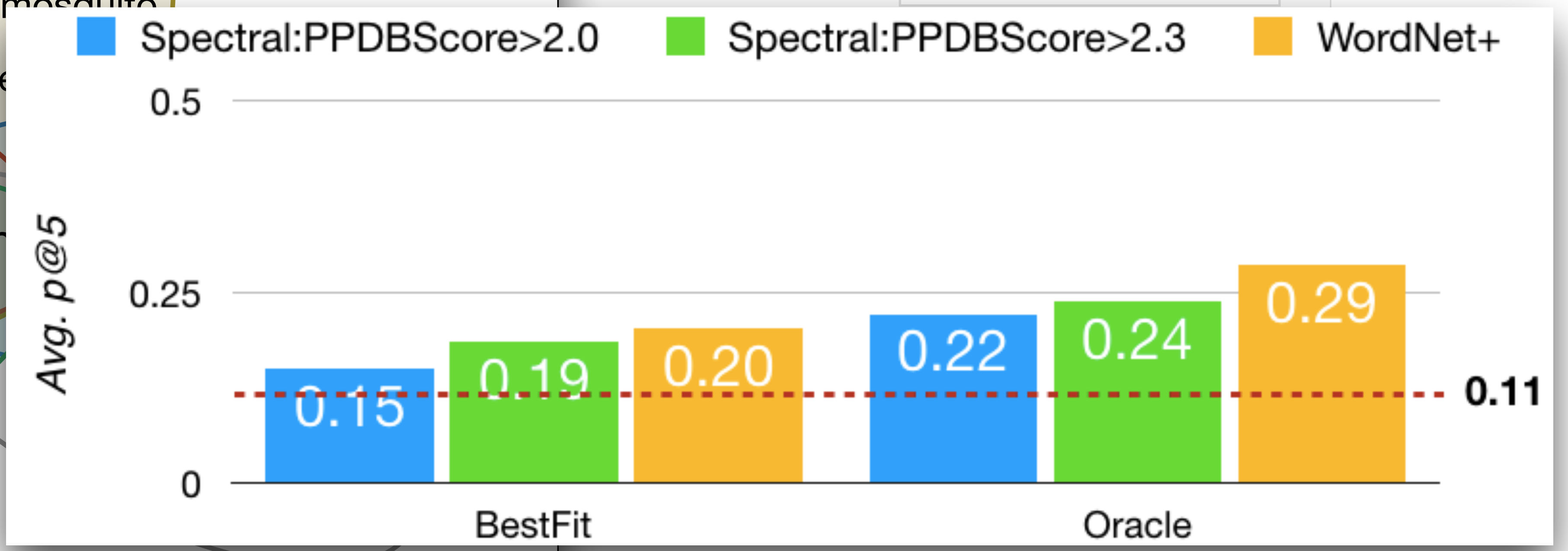
Motivation

Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

sity





Motivation

Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

Learning Scalar Adjective Intensity

EMNLP 2018

Meaning-specific Examples of Word Use

In submission

Conclusion & Future Work

Motivation

Paraphrase pair…			…is evidence that
particularly pleased	⟷	*ecstatic*	*pleased* < *ecstatic*
quite limited	⟷	*restricted*	*limited* < *restricted*
rather odd	⟷	*crazy*	*odd* < *crazy*
so silly	⟷	*dumb*	*silly* < *dumb*
completely mad	⟷	*crazy*	*mad* < *crazy*
$RB\ JJ_1$	⟷	JJ_2	$JJ_1 < JJ_2$

intensifying adverb

Conclusion & Future Work

Motivation

Paraphrase pair ... is evidence that

particularly ... ecstatic

quite lin... estricted

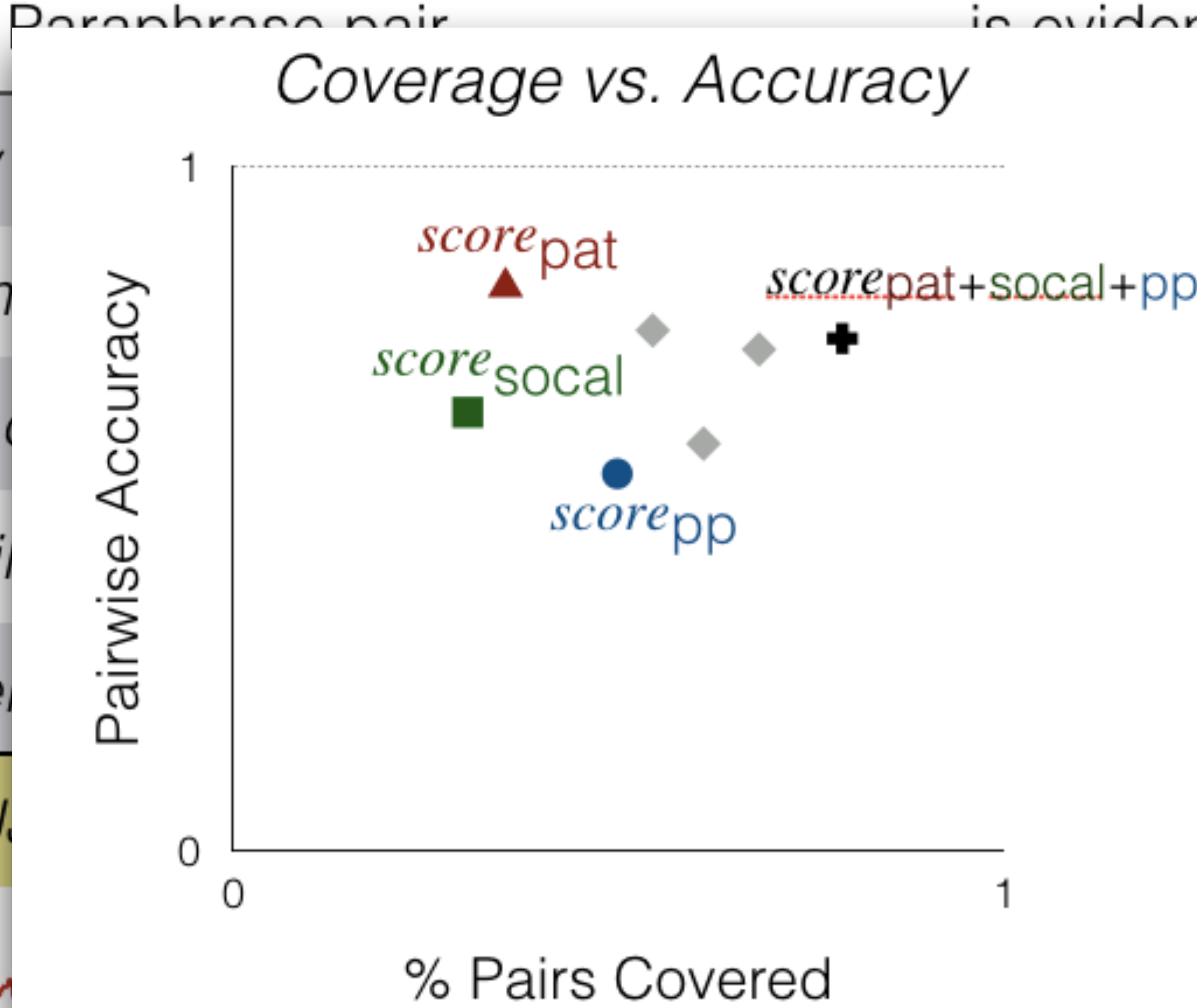
rather ... crazy

so si... lumb

complete... crazy

RB J... JJ_2

inte...



Conclusion & Future Work

Motivation

Using Paraphrases to Model Word Sense

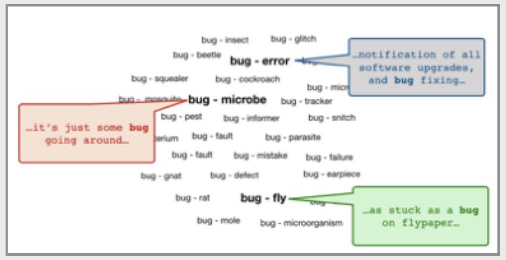
NAACL 2016; SENSE@EACL 2017

Learning Scalar Adjective Intensity

EMNLP 2018

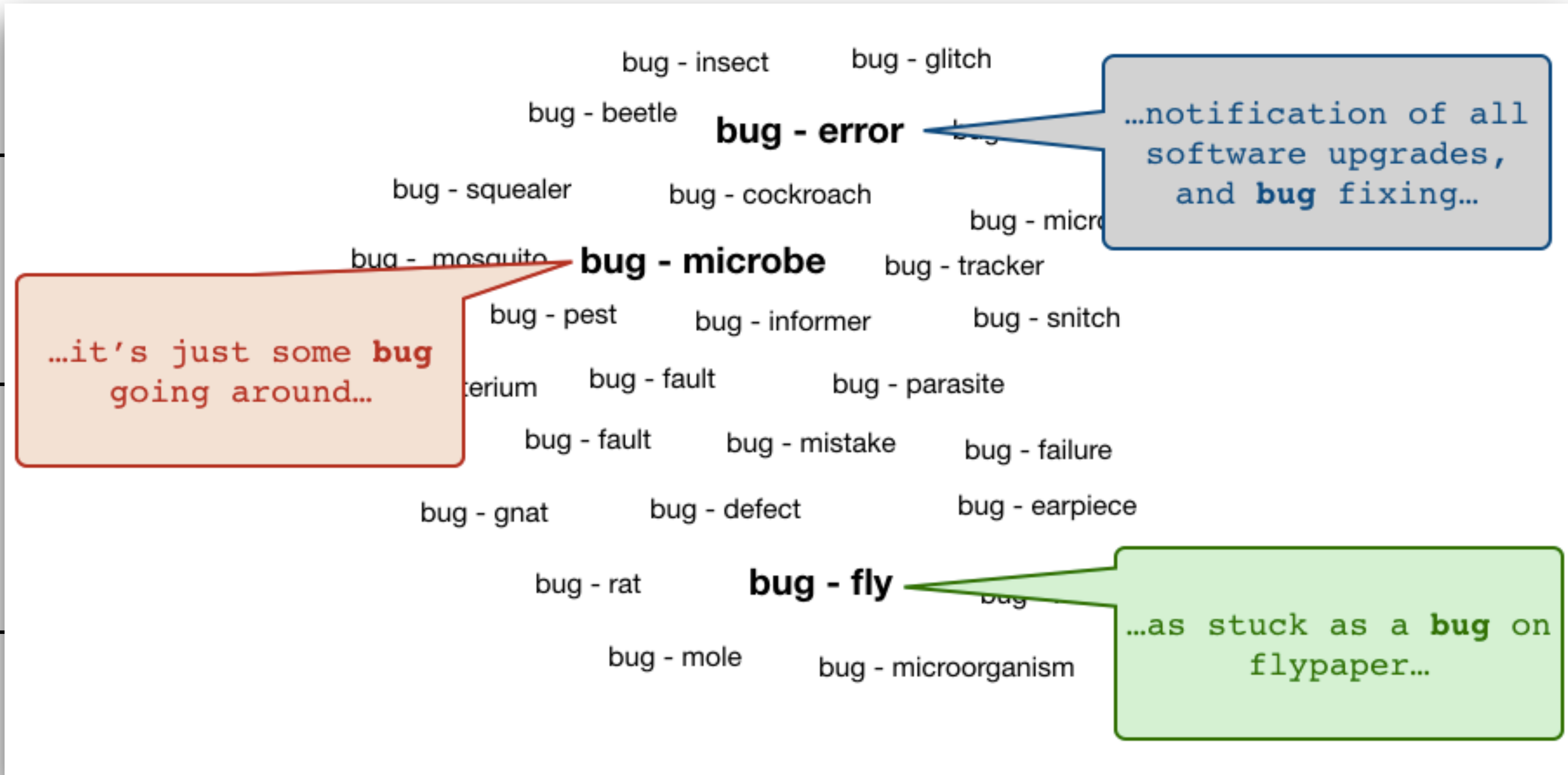
Meaning-specific Examples of Word Use

In submission



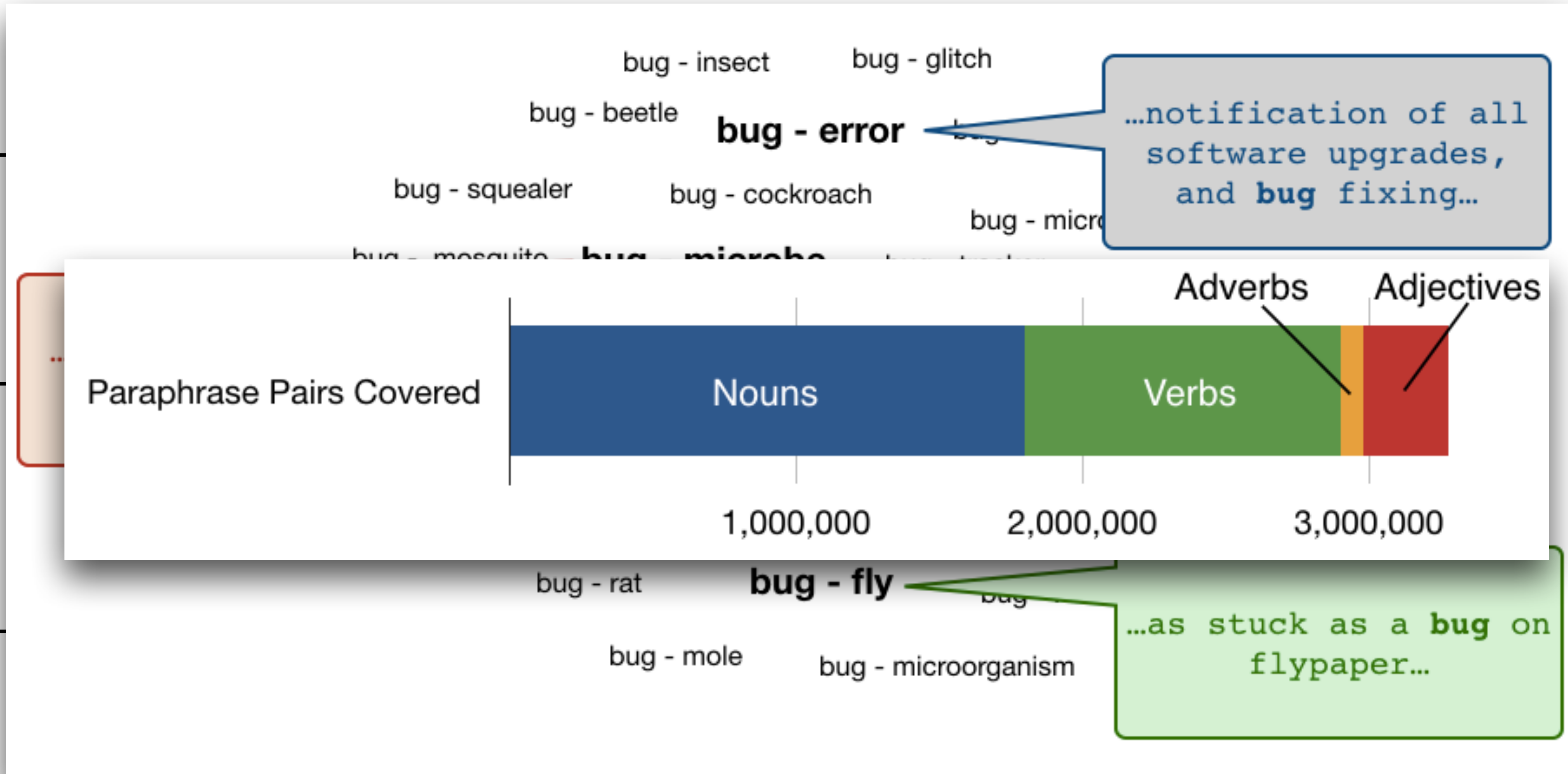
Conclusion & Future Work

Motivation



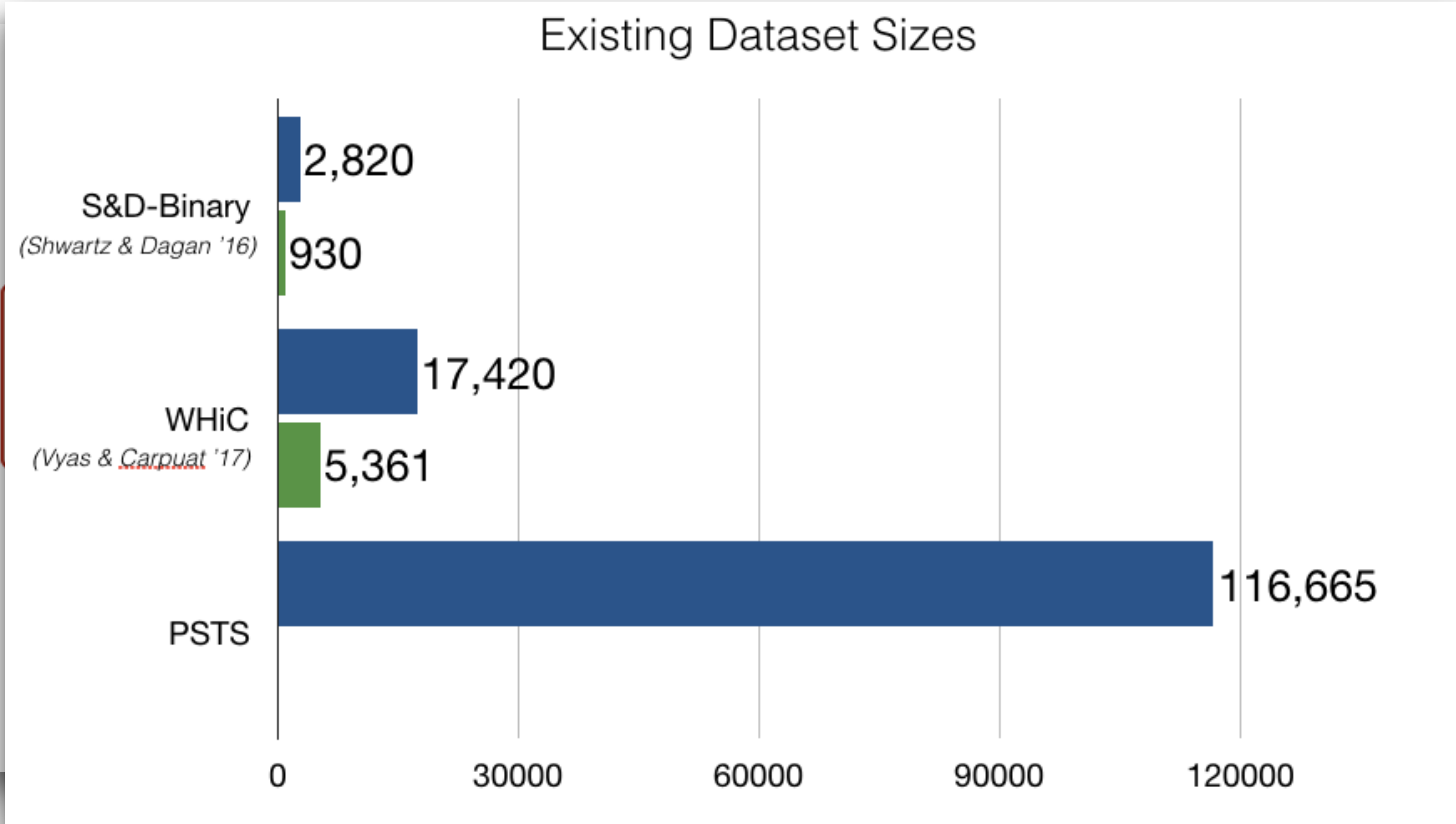
Conclusion & Future Work

Motivation



Conclusion & Future Work

Motivation



Conclusion & Future Work

Motivation

Using Paraphrases to Model Word Sense

NAACL 2016; SENSE@EACL 2017

Learning Scalar Adjective Intensity

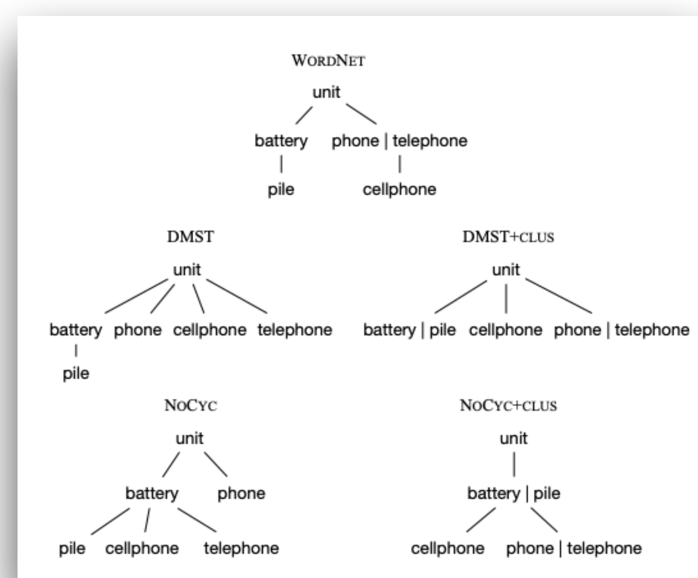
EMNLP 2018

Meaning-specific Examples of Word Use

In submission

Conclusion & Future Work

Future work:
Applying paraphrases to add’l models of lexical semantics



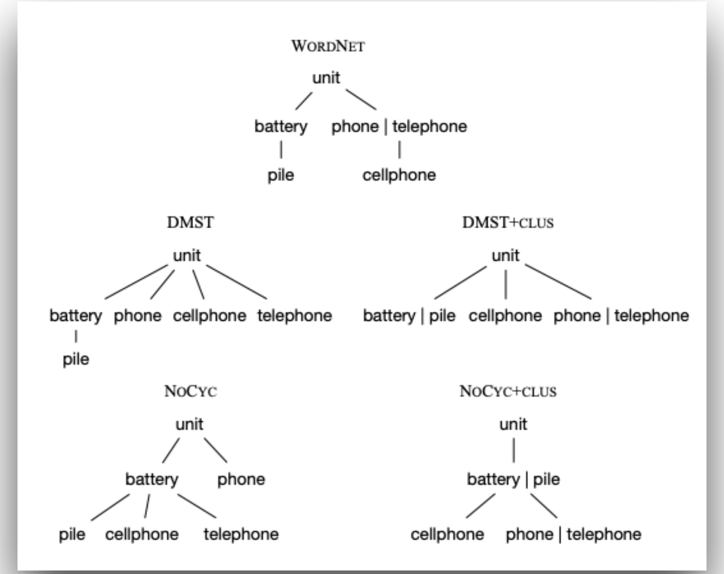
Taxonomy/Ontology Induction?

puppy <--> small dog

Hypernym prediction?

Future work:
Applying paraphrases to add'l models of lexical semantics

- Ripe areas:
  - Require awareness of word sense
  - Benefit from high-coverage features
  - Can learn from comparing phrases to single words



Taxonomy/Ontology Induction?

puppy <--> small dog

Hypernym prediction?

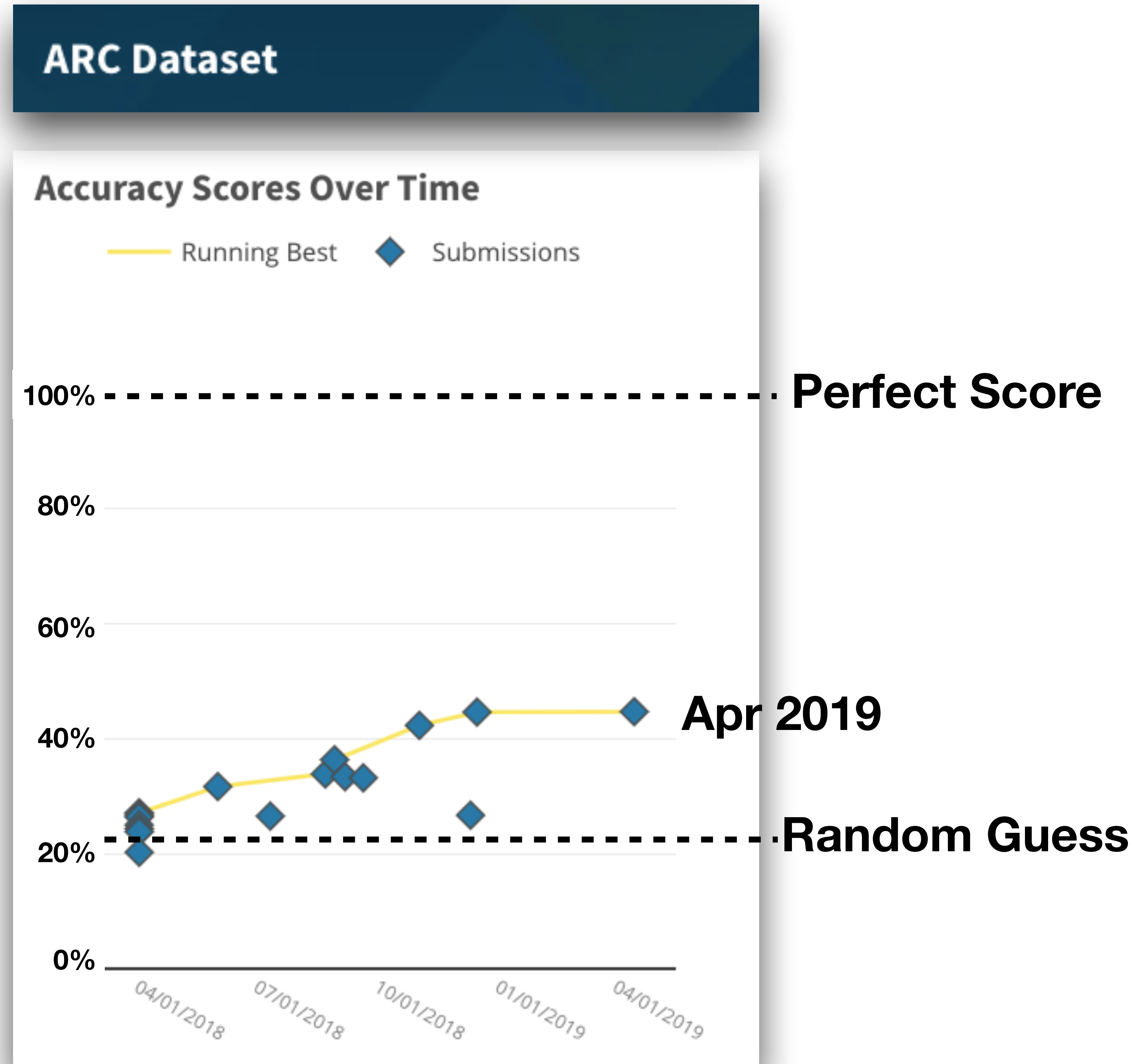
Future work: Integrating lexical semantic knowledge into end-to-end models for downstream tasks

SQuAD2.0

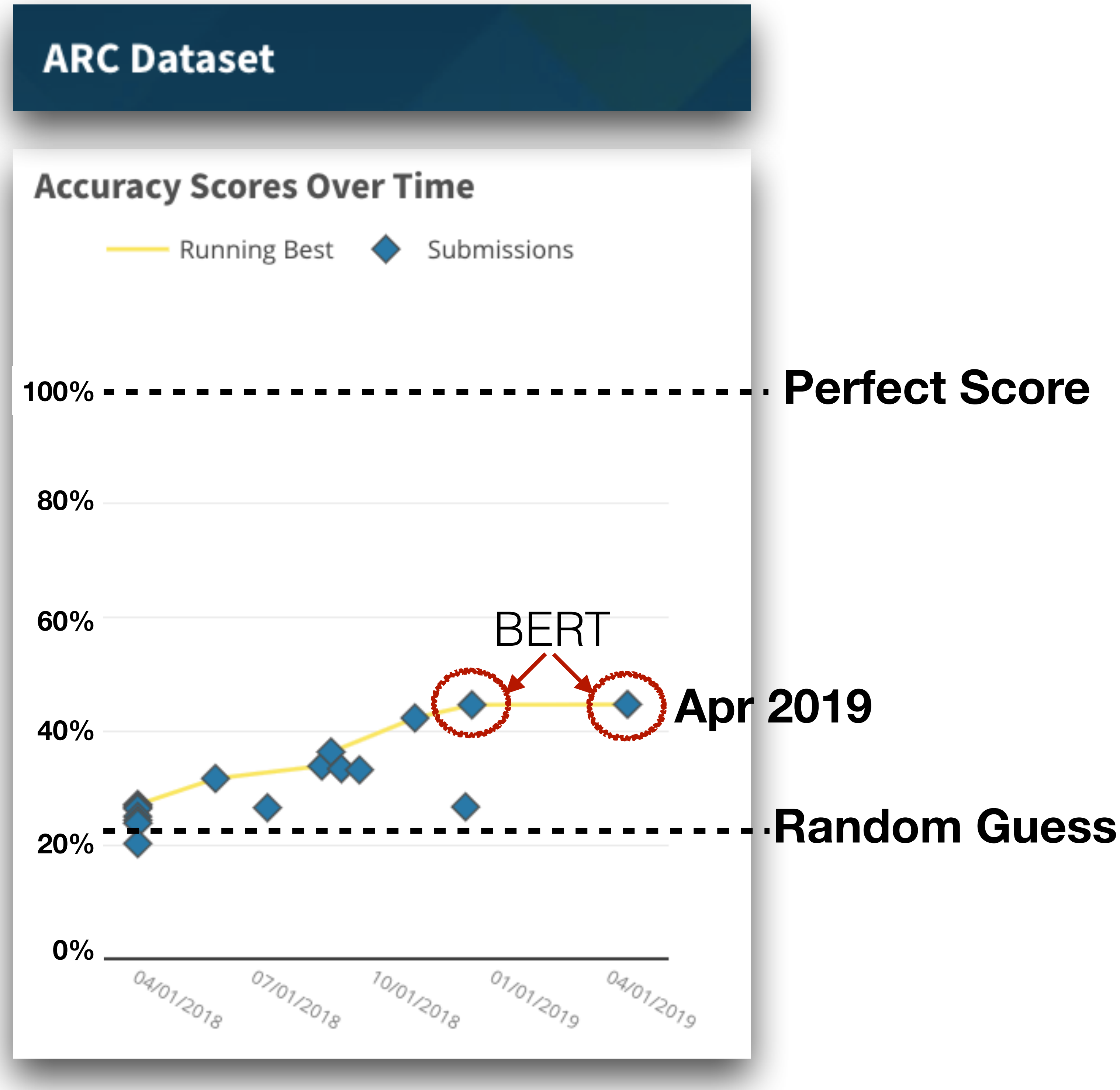
The Stanford Question Answering Dataset

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147

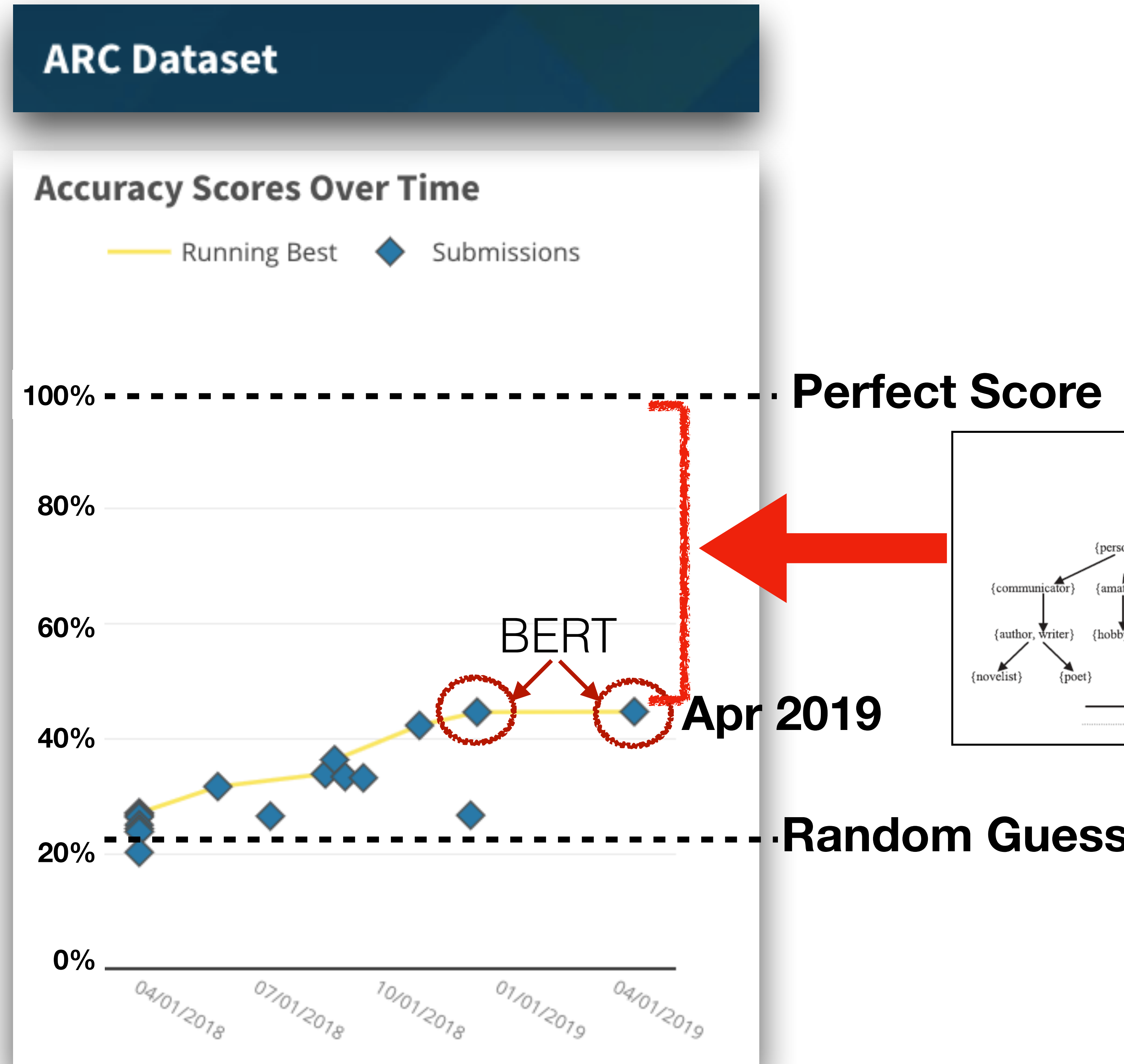
Future work: Integrating lexical semantic knowledge into end-to-end models for downstream tasks

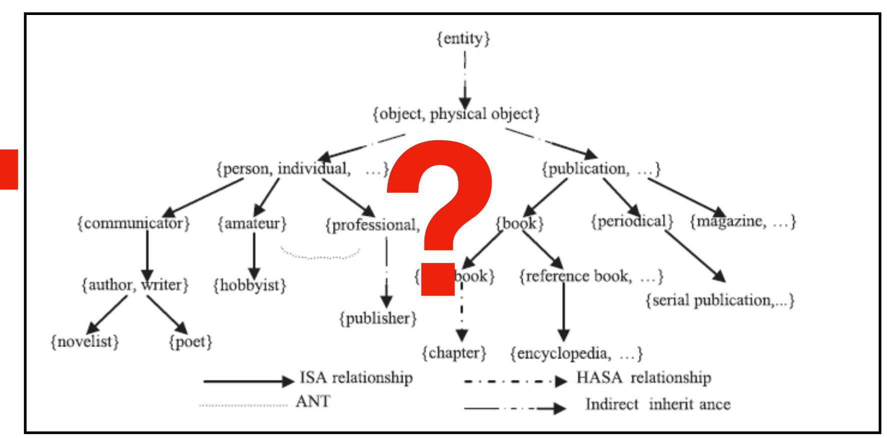


Future work: Integrating lexical semantic knowledge into end-to-end models for downstream tasks



Future work: Integrating lexical semantic knowledge into end-to-end models for downstream tasks





Thank you!

thank you for your time

many thanks

anyway , thanks

here you go

leave a message

gee , thanks

thanks , man

you look amazing

bless you

Thank you!

thank you very much

thank you for your attention

keep the change

uh , thanks

why , thank you

don't thank me

hey , thanks

thank you , frank