Current Semantic-change Quantification Methods Struggle with Discovery in the Wild

Khonzoda Umarova, Lillian Lee, Laerdon Kim
Cornell University

EMNLP 2025

Presenter: Khonzoda Umarova
Contact: ku47@cornell.edu

code/data

Change in the meaning of words over time

The section of a cylinder cut by any ***plane*** *inclined to its axis is an ellipsis*

(19th century)

Schlechtweg et al. (2020)

Change in the meaning of words over time

The section of a cylinder cut by any ***plane*** *inclined to its axis is an ellipsis*

(19th century)

What the hell is you were supposed to be on the ***plane*** *fifteen minutes ago*

(20th century)

Schlechtweg et al. (2020)

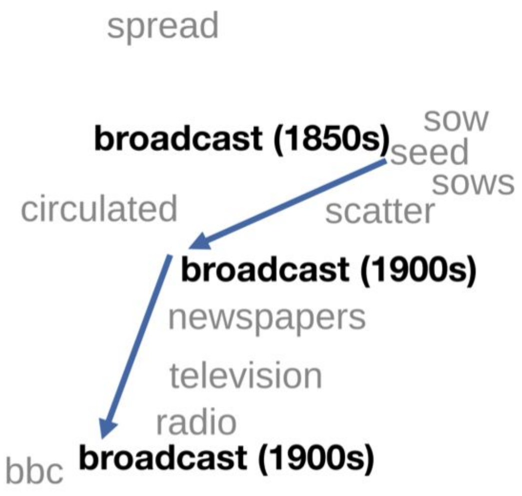
Change in the meaning of words over time

The section of a cylinder cut by any ***plane*** *inclined to its axis is an ellipsis*

(19th century)

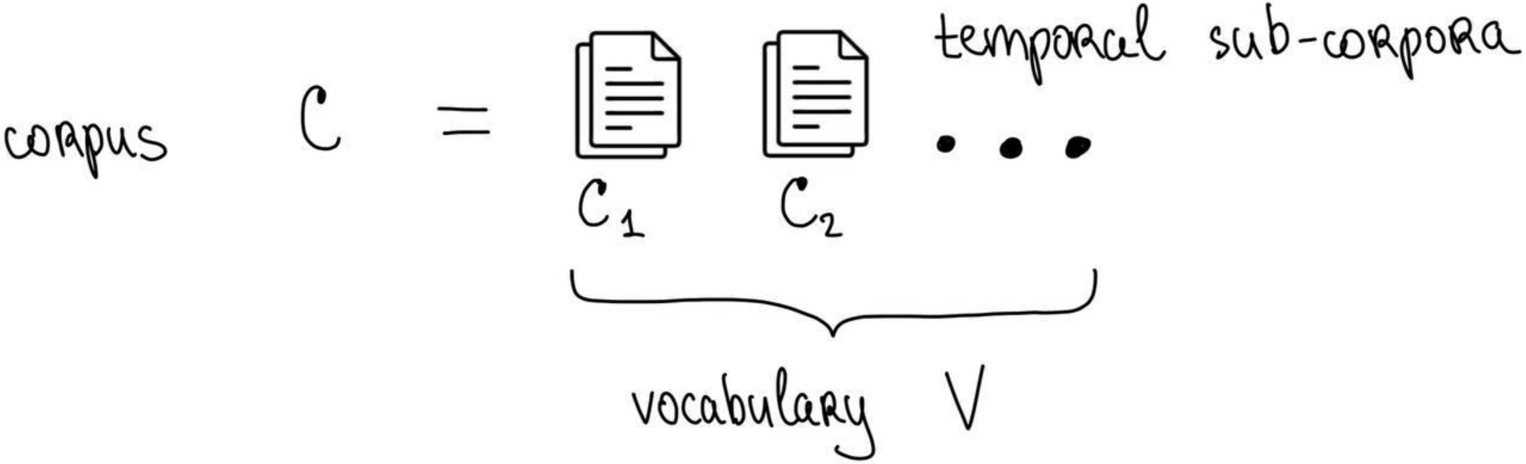
What the hell is you were supposed to be on the ***plane*** *fifteen minutes ago*

(20th century)



Schlechtweg et al. (2020)

Hamilton et al. (2016)



temporal sub-corpora

corpus $C = C_1 \; C_2 \; \dots$

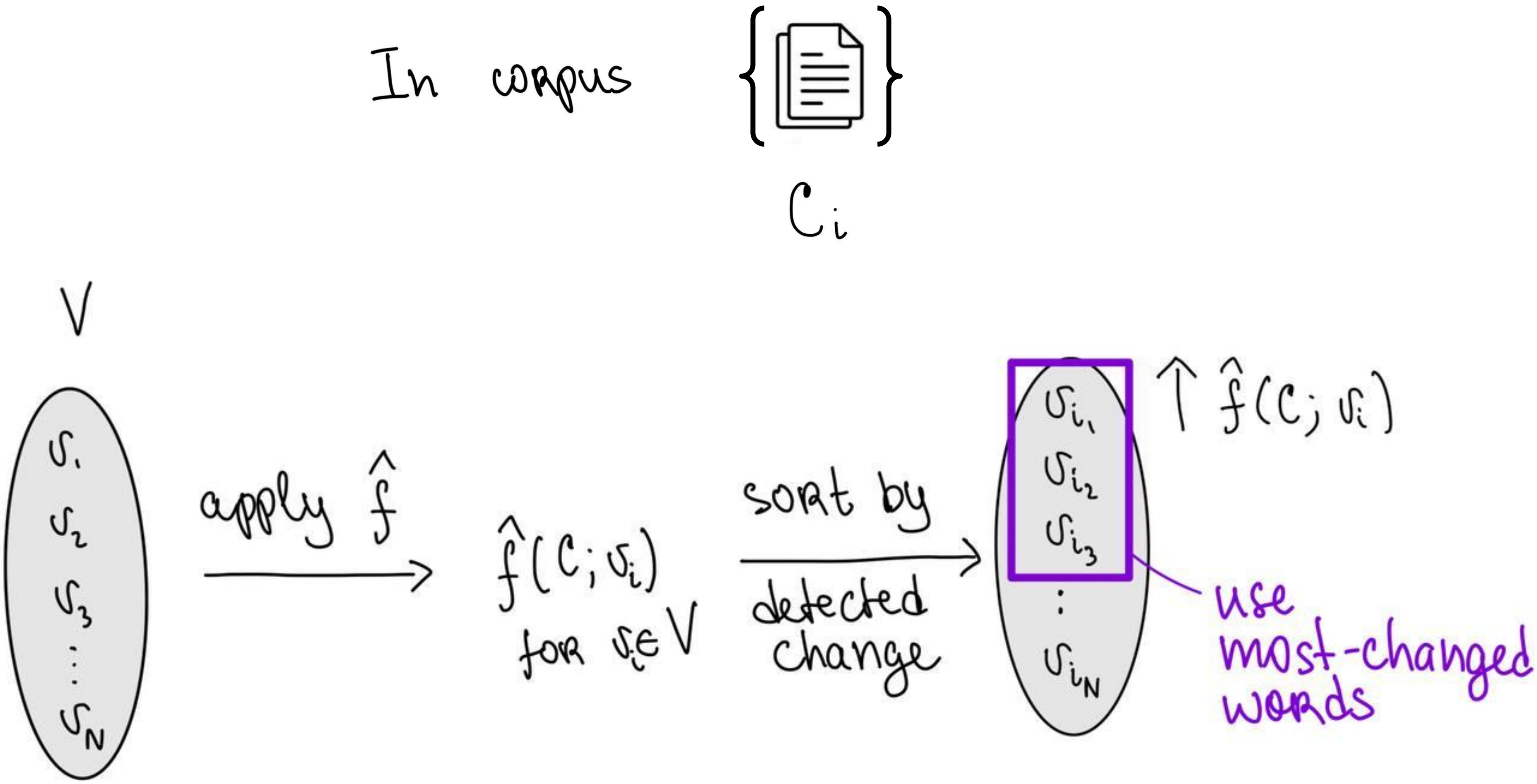
vocabulary V

temporal sub-corpora

corpus $C = \underbrace{C_1 \quad C_2 \quad \dots}_{\text{vocabulary } V}$

semantic-change quantification

scorer $\hat{f} : C, V \rightarrow [0, 1]$


In corpus
C_i
V
v_1
v_2
v_3
$\vdots$
v_N
apply $\hat{f}$
$\hat{f}(C; v_i)$
for $v_i \in V$
sort by
detected change
v_{i_1}
v_{i_2}
v_{i_3}
$\vdots$
v_{i_N}
$\uparrow \hat{f}(C; v_i)$
use most-changed words

Semantic-change detection

Generally semantic-change detection methods are evaluated on a set of benchmarks where

$T \subseteq V$

Schlechtweg et al. (2020), Del Tredici et al. (2019), Kutuzov and Pivovarova (2021), ...

Semantic-change detection

Generally semantic-change detection methods are evaluated on a set of benchmarks where

$T \subseteq V$

↳ annotate $t \in T$: $\ell(t; C)$

Schlechtweg et al. (2020), Del Tredici et al. (2019), Kutuzov and Pivovarova (2021), ...

Semantic-change detection

Generally semantic-change detection methods are evaluated on a set of benchmarks where

$T \subseteq V$

↓ annotate $t \in T$: $\underline{\underline{\ell(t;C)}}$

↳ compare to $\hat{f}(t;C)$

Schlechtweg et al. (2020), Del Tredici et al. (2019), Kutuzov and Pivovarova (2021), …

Semantic-change detection

Generally semantic-change detection methods are evaluated on a set of benchmarks where

$T \subseteq V$

↳ annotate $t \in T$: $\underline{\underline{\ell(t;C)}}$

↳ compare to $\hat{f}(t;C)$

Schlechtweg et al. (2020), Del Tredici et al. (2019), Kutuzov and Pivovarova (2021), ...

However, this evaluation doesn't illustrate discovery performance of $\hat{f}$ outside T

Evaluating semantic-change *discovery*

Without **$\ell(v, C)$** for all **$v \in V$**, it is challenging to evaluate semantic-change discovery.

Evaluating semantic-change *discovery*

Without $\boldsymbol{\ell}(\mathbf{v}, \mathbf{C})$ for all $\mathbf{v} \in \mathbf{V}$, it is challenging to evaluate semantic-change discovery.

Kurtyigit et al. (2021)

Evaluating semantic-change *discovery*

Without **ℓ(v, C)** for all **v∈V**, it is challenging to evaluate semantic-change discovery.

Alternative: two mutually complementary approaches

Evaluating semantic-change *discovery*

Without **$\ell(v, C)$** for all **$v \in V$**, it is challenging to evaluate semantic-change discovery.

Alternative: two mutually complementary approaches

[ranking-based] how well do semantic-change quantification methods ($\hat{f}$) rank known high changes?

Evaluating semantic-change *discovery*

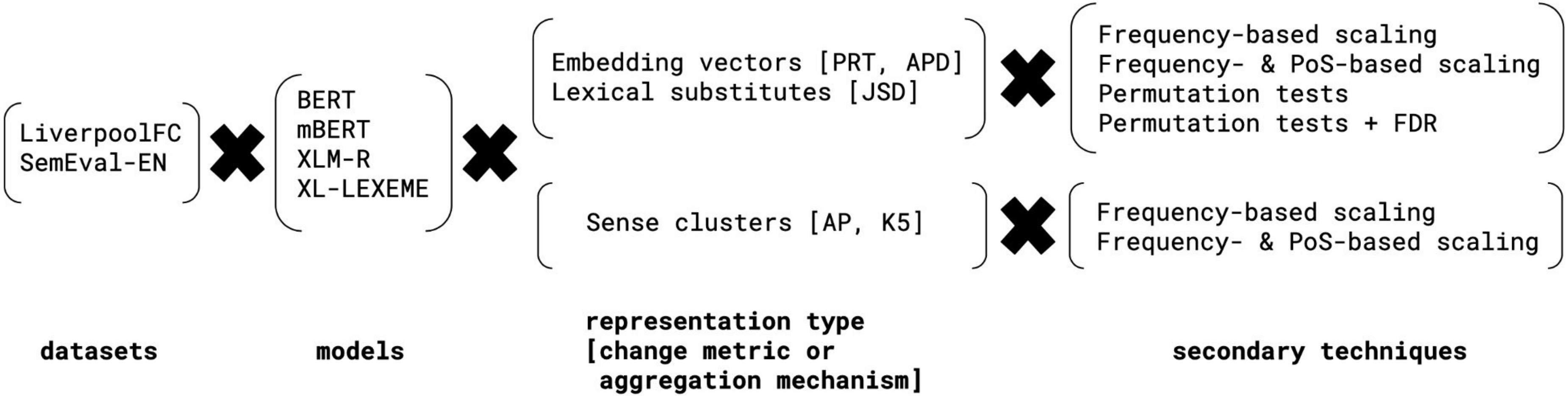
Without **$\ell(v, C)$** for all **$v \in V$**, it is challenging to evaluate semantic-change discovery.

Alternative: two mutually complementary approaches

[ranking-based] how well do semantic-change quantification methods ($\hat{f}$) rank known high changes?

[annotations-based] are terms **$v \in V$** ranked high by semantic-change quantification methods ($\hat{f}$) in fact genuine semantic changes?

Evaluating semantic-change *discovery*



LiverpoolFC
SemEval-EN

×

BERT
mBERT
XLM-R
XL-LEXEME

×

Embedding vectors [PRT, APD]
Lexical substitutes [JSD]

×

Frequency-based scaling
Frequency- & PoS-based scaling
Permutation tests
Permutation tests + FDR

Sense clusters [AP, K5]

×

Frequency-based scaling
Frequency- & PoS-based scaling

datasets

models

representation type [change metric or aggregation mechanism]

secondary techniques

Ranking-based evaluation

Ranking-based evaluation

Within T (annotated subset of V) consider $T^* = \{t \in T \text{ such that } \ell(C, t) > \beta\}$

T* represents **highest known** semantic changes

Ranking-based evaluation

Within T (annotated subset of V) consider $T^* = \{t \in T \text{ such that } \ell(C, t) > \beta\}$

T* represents **highest known** semantic changes

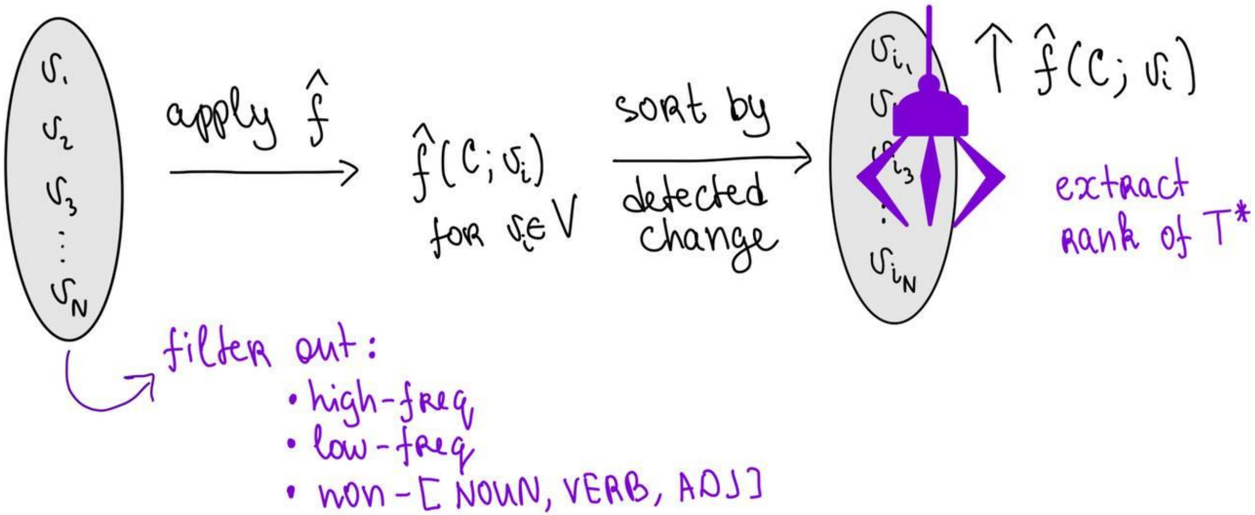
Hence, semantic-change detection methods ***should*** rank T* well

Ranking-based evaluation

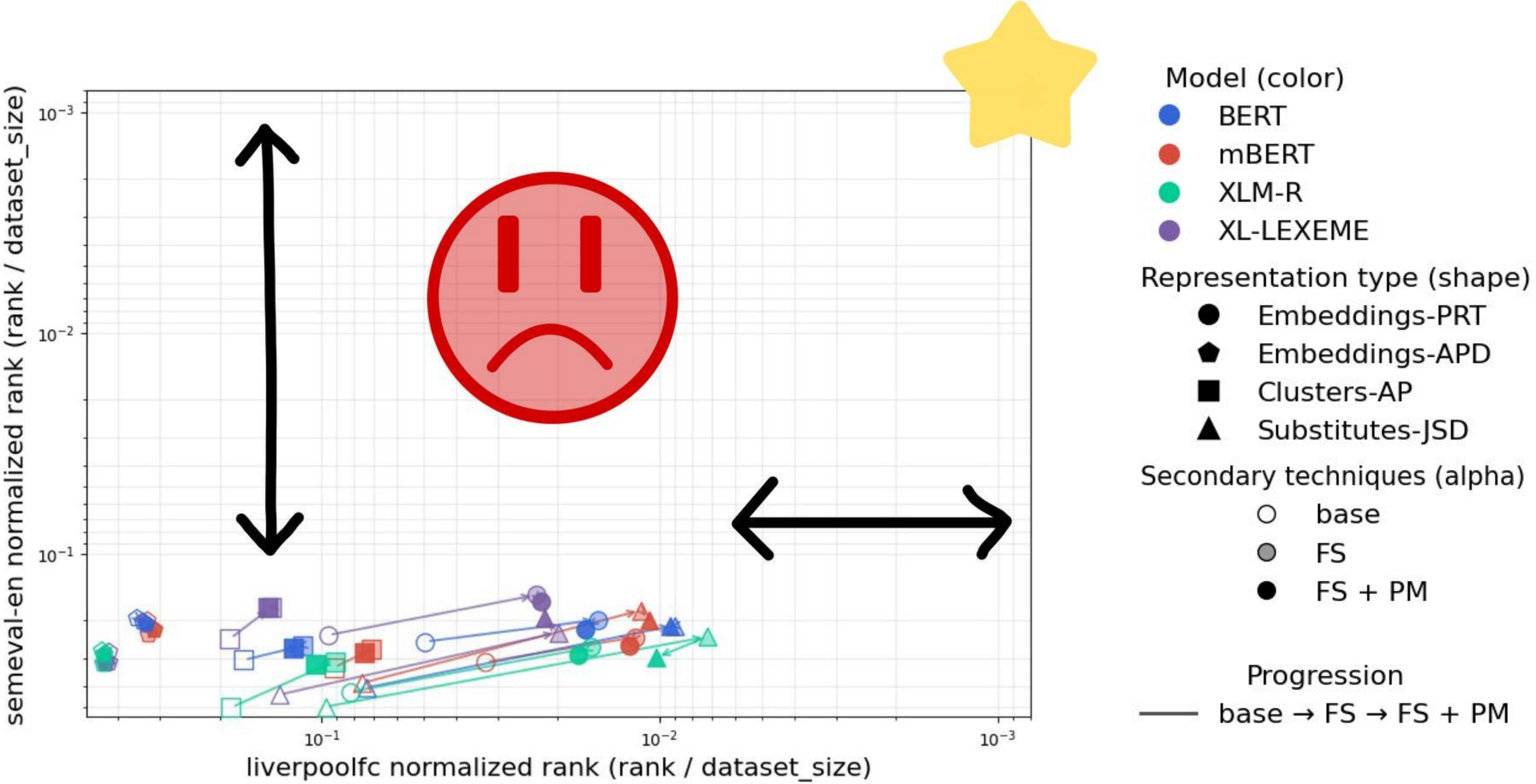
Within T (annotated subset of V) consider $T^* = \{t \in T \text{ such that } \ell(C, t) > \beta\}$

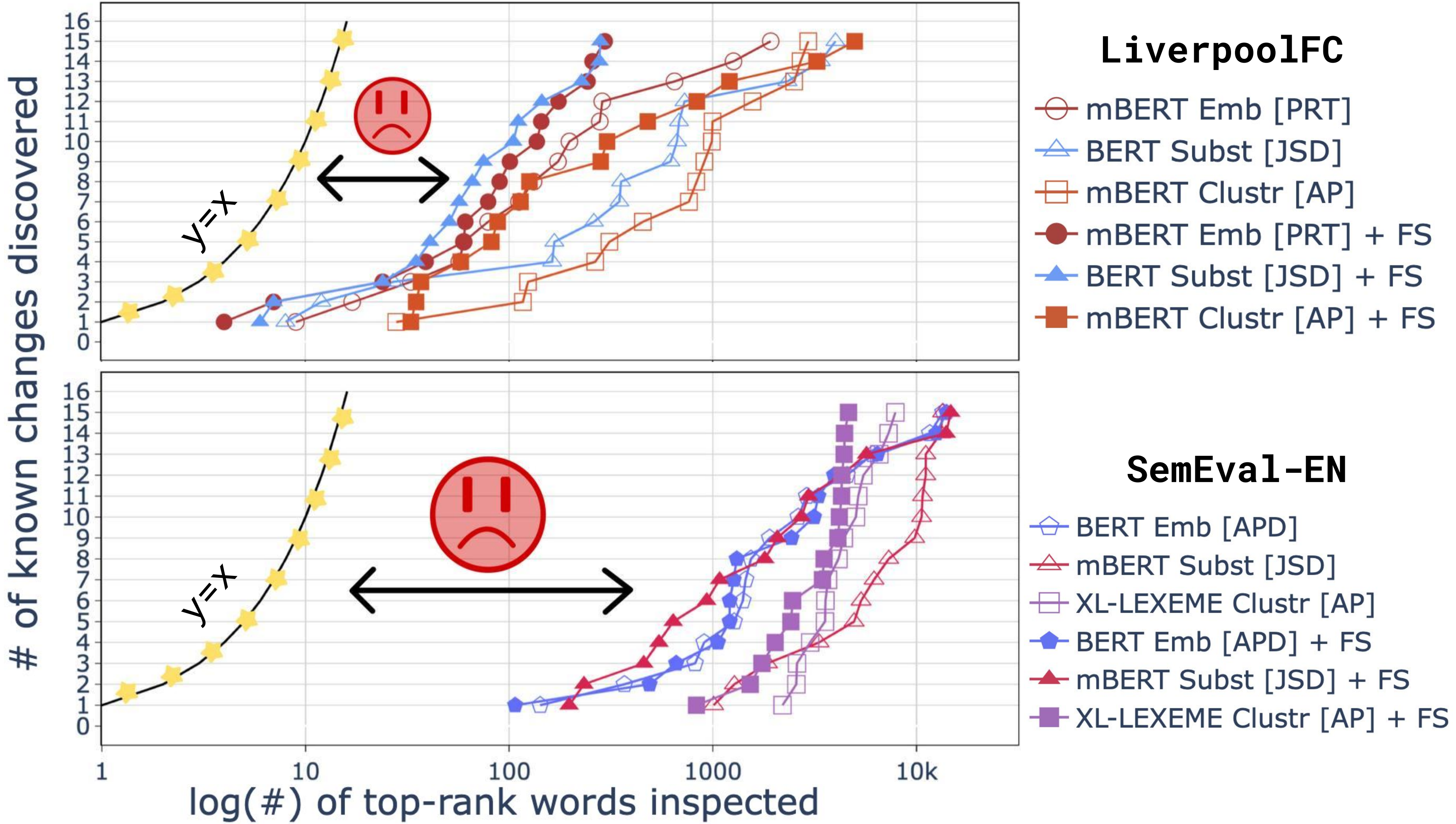
T* represents **highest known** semantic changes

Hence, semantic-change detection methods ***should*** rank T* well



For both datasets: semantic-change quantification methods struggle at ranking T* high




LiverpoolFC
mBERT Emb [PRT]
BERT Subst [JSD]
mBERT Clustr [AP]
mBERT Emb [PRT] + FS
BERT Subst [JSD] + FS
mBERT Clustr [AP] + FS
SemEval-EN
BERT Emb [APD]
mBERT Subst [JSD]
XL-LEXEME Clustr [AP]
BERT Emb [APD] + FS
mBERT Subst [JSD] + FS
XL-LEXEME Clustr [AP] + FS
y=x
y=x
# of known changes discovered
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
log(#) of top-rank words inspected
1
10
100
1000
10k

Annotations-based evaluation 🔍

Ranking of T* alone **doesn't give** full understanding about ***discovery***.

Annotations-based evaluation

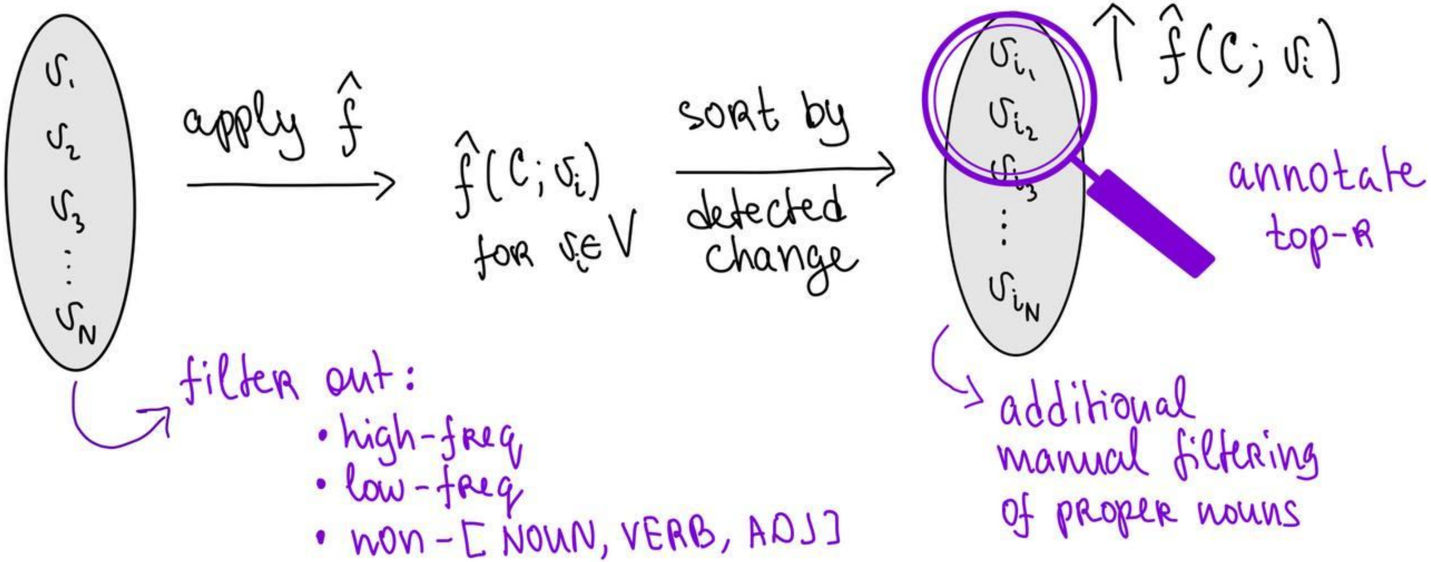
Ranking of T* alone **doesn't give** full understanding about ***discovery***.

If not T*, then what are the words ranked high by $\hat{f}$? Did they change in meaning?

Annotations-based evaluation

Ranking of T* alone **doesn't give** full understanding about ***discovery***.

If not T*, then what are the words ranked high by $\hat{f}$? Did they change in meaning?



Annotations-based evaluation

We select k=15 and annotate top–k-ranked changes (Del Tredici et al., 2019)

Annotations-based evaluation

We select k=15 and annotate top–k-ranked changes (Del Tredici et al., 2019)

GROUP 1

[S1] *you will observe that the two circles partly described on the plate for the purpose of determining the opening of the box are circles standing upright on the ground* ***plane*** *and not laying on it as are the circles in plate 10*

[S2] …

[S3] …

[S4] …

[S5] *the section of a cylinder cut by any* ***plane*** *inclined to its axis is an ellipsis*

GROUP 2

[S1] *chief of staff and his* ***plane*** *lands in two hours*

[S2] …

[S3] …

[S4] …

[S5] *what the hell is you were supposed to be on the* ***plane*** *fifteen minutes ago*

Annotations-based evaluation

We select k=15 and annotate top–k-ranked changes (Del Tredici et al., 2019)

Annotators are asked:

(a) Does group 1 have a majority sense? Group 2?

GROUP 1

[S1] *you will observe that the two circles partly described on the plate for the purpose of determining the opening of the box are circles standing upright on the ground* ***plane*** *and not laying on it as are the circles in plate 10*

[S2] …

[S3] …

[S4] …

[S5] *the section of a cylinder cut by any* ***plane*** *inclined to its axis is an ellipsis*

GROUP 2

[S1] *chief of staff and his* ***plane*** *lands in two hours*

[S2] …

[S3] …

[S4] …

[S5] *what the hell is you were supposed to be on the* ***plane*** *fifteen minutes ago*

Annotations-based evaluation

We select k=15 and annotate top–k-ranked changes (Del Tredici et al., 2019)

Annotators are asked:

(a) Does group 1 have a majority sense? Group 2?

(b) Is the majority sense in group 1 different from that in group 2?

GROUP 1	GROUP 2
[S1] *you will observe that the two circles partly described on the plate for the purpose of determining the opening of the box are circles standing upright on the ground* ***plane*** *and not laying on it as are the circles in plate 10*	**[S1]** *chief of staff and his* ***plane*** *lands in two hours*
[S2] …	**[S2]** …
[S3] …	**[S3]** …
[S4] …	**[S4]** …
[S5] *the section of a cylinder cut by any* ***plane*** *inclined to its axis is an ellipsis*	**[S5]** *what the hell is you were supposed to be on the* ***plane*** *fifteen minutes ago*

Annotations-based evaluation

We select k=15 and annotate top–k-ranked changes (Del Tredici et al., 2019)

Annotators are asked:

(a) Does group 1 have a majority sense? Group 2?

(b) Is the majority sense in group 1 different from that in group 2?

(c) What are the sentences whose senses appear in group 1 but not in group 2 (and vice versa)?

GROUP 1	GROUP 2
[S1] *you will observe that the two circles partly described on the plate for the purpose of determining the opening of the box are circles standing upright on the ground **plane** and not laying on it as are the circles in plate 10* **[S2]** … **[S3]** … **[S4]** … **[S5]** *the section of a cylinder cut by any **plane** inclined to its axis is an ellipsis*	**[S1]** *chief of staff and his **plane** lands in two hours* **[S2]** … **[S3]** … **[S4]** … **[S5]** *what the hell is you were supposed to be on the **plane** fifteen minutes ago*

Annotations-based evaluation

We select k=15 and annotate top–k-ranked changes (Del Tredici et al., 2019)

We recruit English speakers who annotate

- 76+ words in SemEval-EN and
- 83+ words in LiverpoolFC

for semantic change using this approach

Annotations-based evaluation

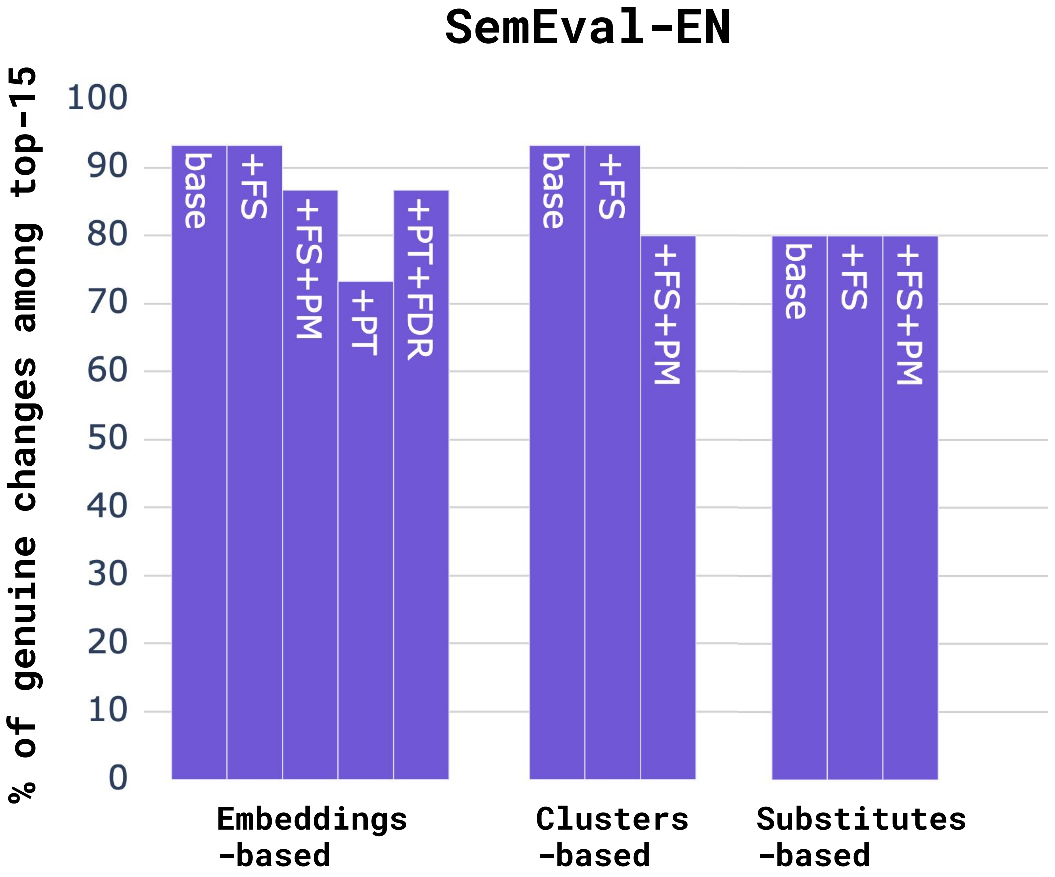
We select k=15 and annotate top–k-ranked changes (Del Tredici et al., 2019)

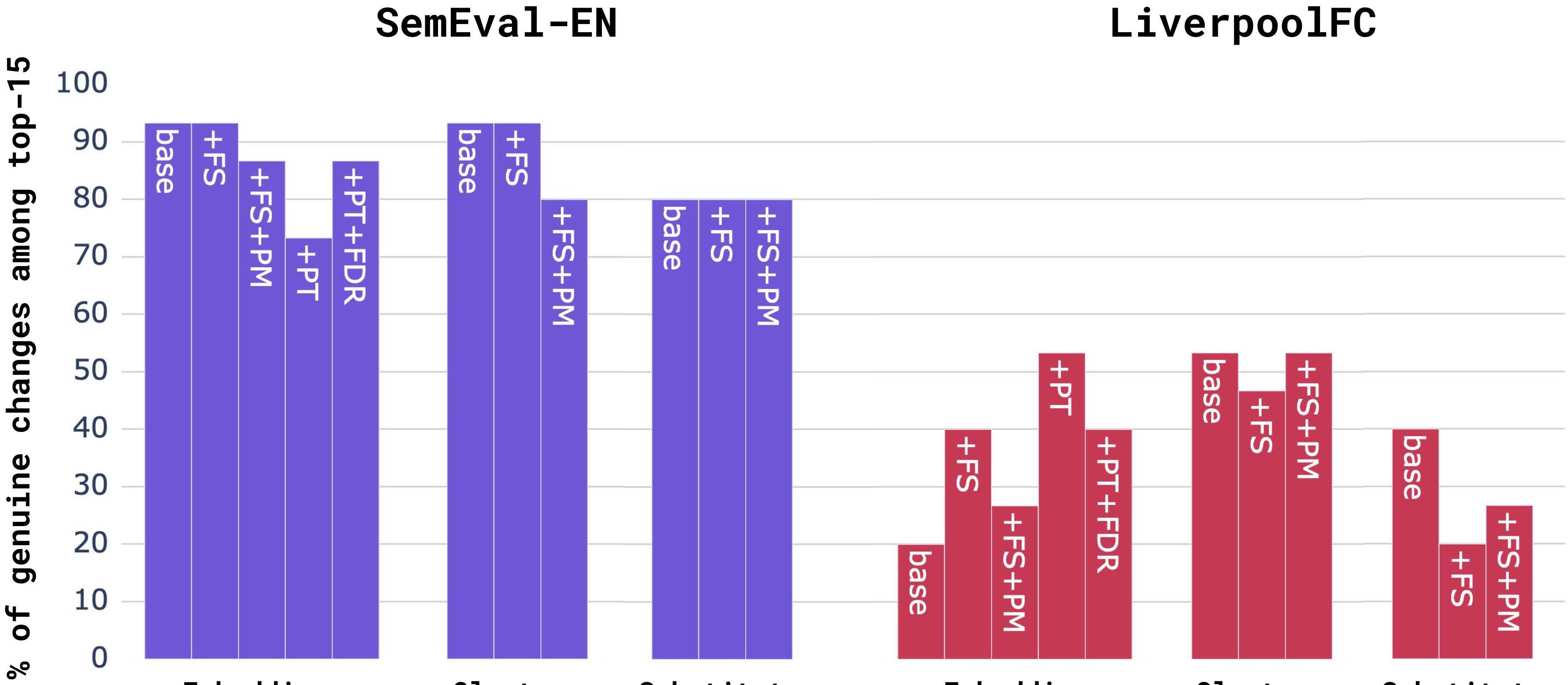
We recruit English speakers who annotate

- 76+ words in SemEval-EN and
- 83+ words in LiverpoolFC

for semantic change using this approach

annotations


SemEval-EN
% of genuine changes among top-15
100
90
80
70
60
50
40
30
20
10
0
base
+FS
+FS+PM
+PT
+PT+FDR
base
+FS
+FS+PM
base
+FS
+FS+PM
Embeddings -based
Clusters -based
Substitutes -based


SemEval-EN
LiverpoolFC
% of genuine changes among top-15
100
90
80
70
60
50
40
30
20
10
0
base
+FS
+FS+PM
+PT
+PT+FDR
base
+FS
+FS+PM
base
+FS
+FS+PM
base
+FS
+FS+PM
+PT
+PT+FDR
base
+FS
+FS+PM
base
+FS
+FS+PM
Embeddings -based
Clusters -based
Substitutes -based
Embeddings -based
Clusters -based
Substitutes -based

Discussion

Short-term semantic change in LiverpoolFC

Discussion

Short-term semantic change in LiverpoolFC

- "Data bursts" (Kutuzov et al., 2022)

Discussion

Short-term semantic change in LiverpoolFC

- "Data bursts" (Kutuzov et al., 2022)
- need to excel at ***quantifying*** and ***discovering*** short-term semantic change

Discussion

Short-term semantic change in LiverpoolFC

- "Data bursts" (Kutuzov et al., 2022)
- need to excel at ***quantifying*** and ***discovering*** short-term semantic change

≈ recall-like metric

≈ precision-like metric

two different perspectives, but together they give a better understanding of **semantic-change discovery**

Thank you!

Contact: **ku47@cornell.edu**

code/data: