SEOUL NATIONAL UNIV. [1]
VISION & LEARNING

Code available at
github.com/seilna/cnn-units-in-nlp

Discovery of Natural Language Concepts in Individual Units of CNNs

Seil Na[1]
Seoul National University

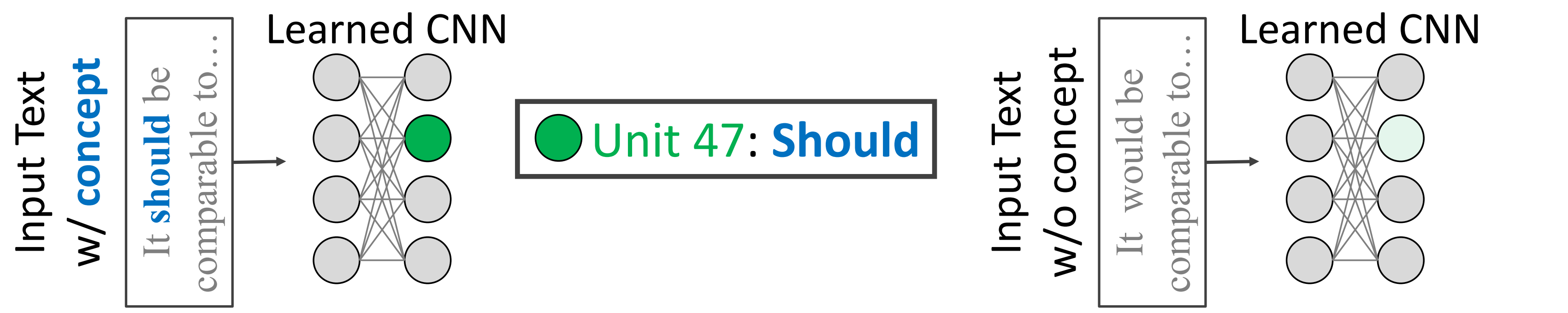
Yo Joong Choe[2]
Kakao

Dong-Hyun Lee[3]
Kakao Brain

Gunhee Kim[1]
Seoul National University

kakao[2]
kakaobrain[3]

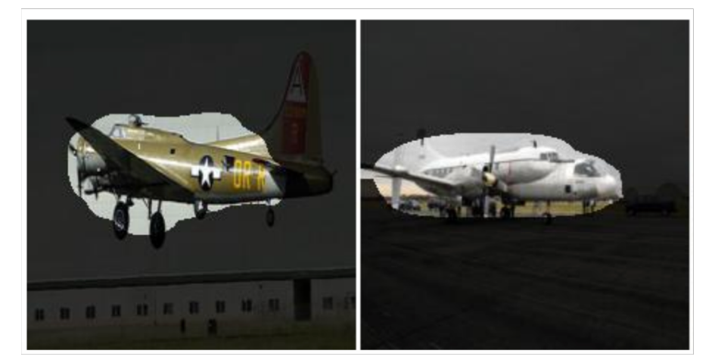
Goal: Unit-level Analysis of Natural Language Representation



Contribution We show that **individual units** of CNNs learned on NLP tasks could act as **natural language concept** detectors

Why Unit-level Analysis of Representation?

Unit 151: **airplane**



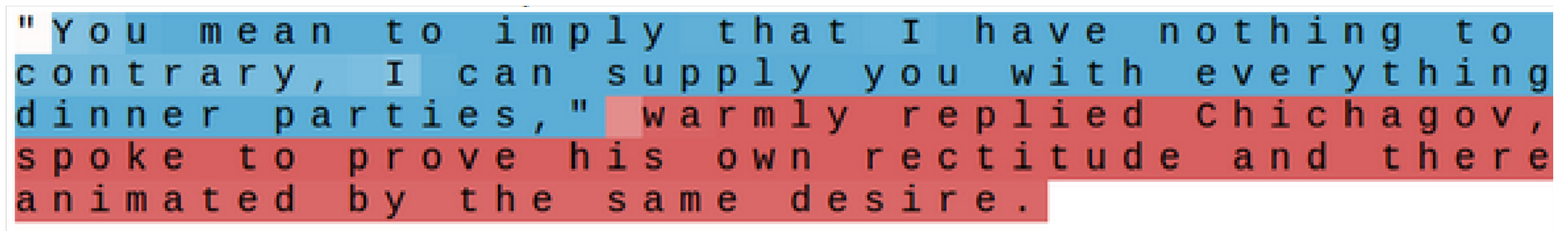
More fine-grained insights of representation (Bau et al,. CVPR 17')

living room(0.31) sofa(37.87%) cabinet(12.32%)

Generate explanation of given prediction (Zhou et al., ECCV 18')

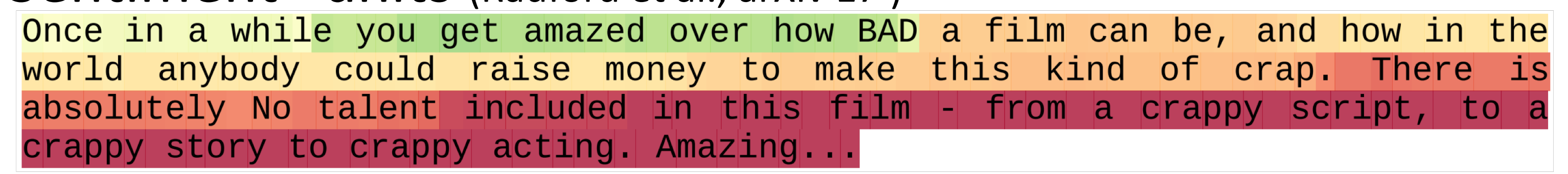
Previous Work on Unit-level Analysis

- "Quote" units (Karpathy et al., ICLR workshop 16')



```
"You mean to imply that I have nothing to
contrary, I can supply you with everything
dinner parties," warmly replied Chichagov,
spoke to prove his own rectitude and there
animated by the same desire.
```

- "Sentiment" units (Radford et al., arXiv 17')



```
Once in a while you get amazed over how BAD a film can be, and how in the
world anybody could raise money to make this kind of crap. There is
absolutely No talent included in this film - from a crappy script, to a
crappy story to crappy acting. Amazing...
```

- "**Natural Language Concept***" units (**Ours**)

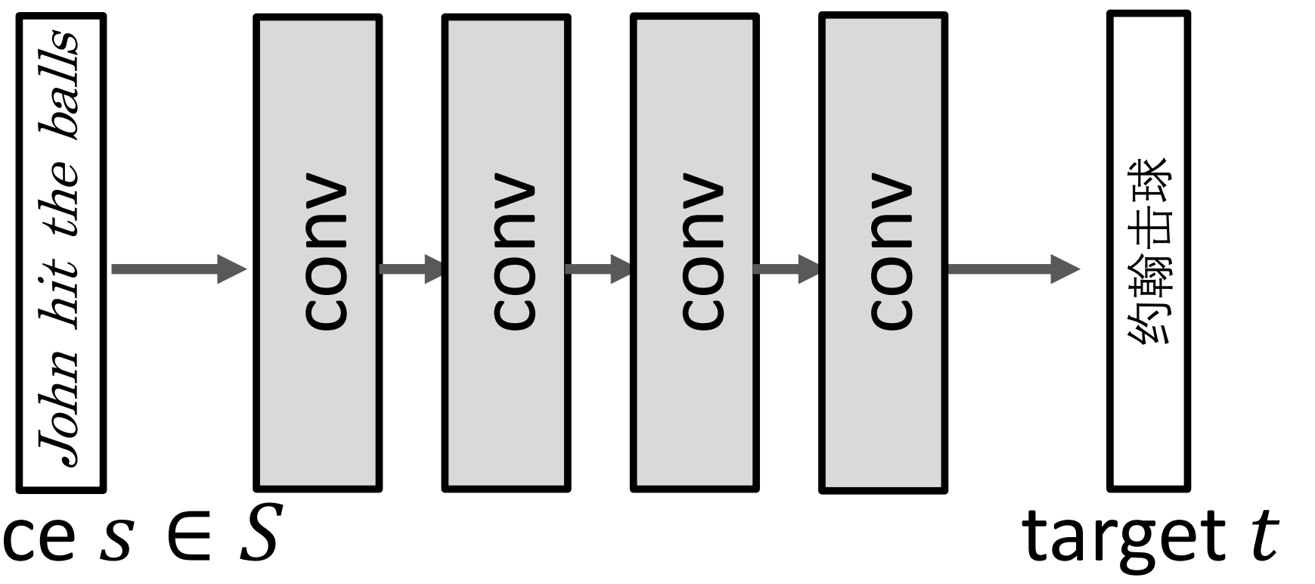
Unit 711: *should*, *would*, *not*, *can*

- That would not be democratic.
- That would be cheap and it would not be right.
- This is not how it should be in a democracy.
- I hope that you would not want that!
- Europe cannot and must not tolerate this.

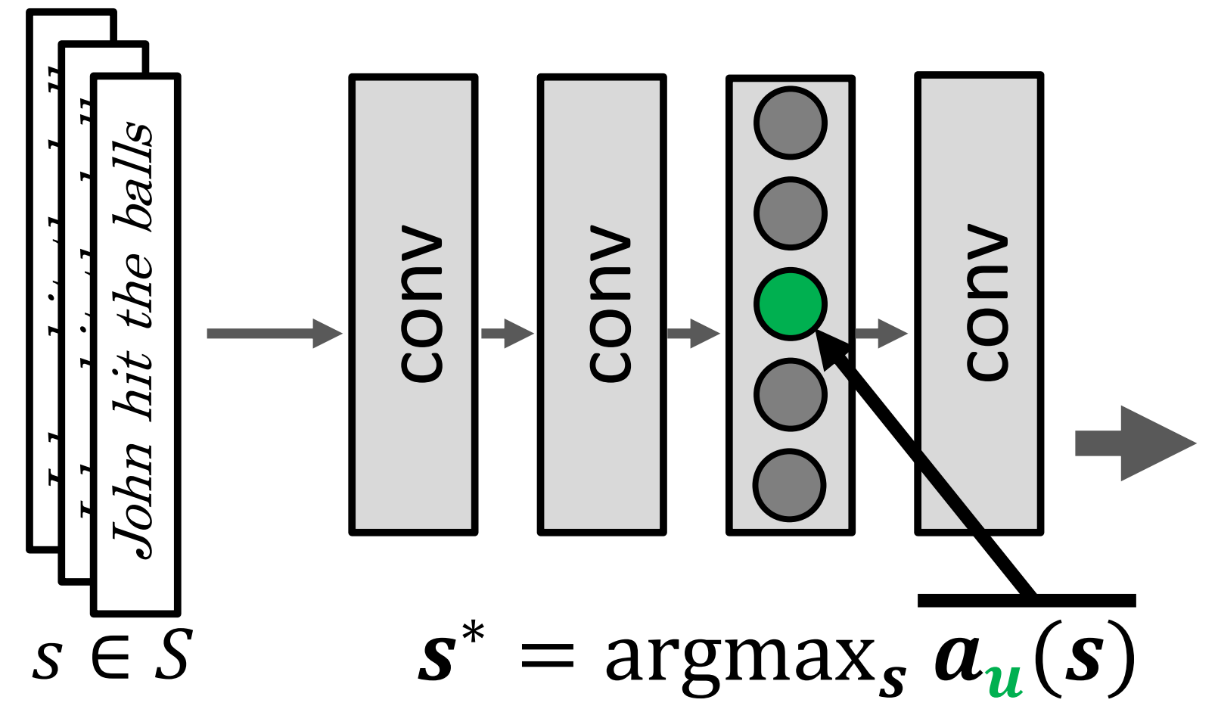
*We define concept as building blocks of natural language sentence; [**Morpheme** / **Word** / **Phrase**]

Approach: Alignment Score between Units and Concepts

1. Train CNN (e.g. ByteNet) on language task (e.g. Translation)



2. For each unit $\boldsymbol{u}$, find **top** $\boldsymbol{k}$ **sentences** which highly activate it

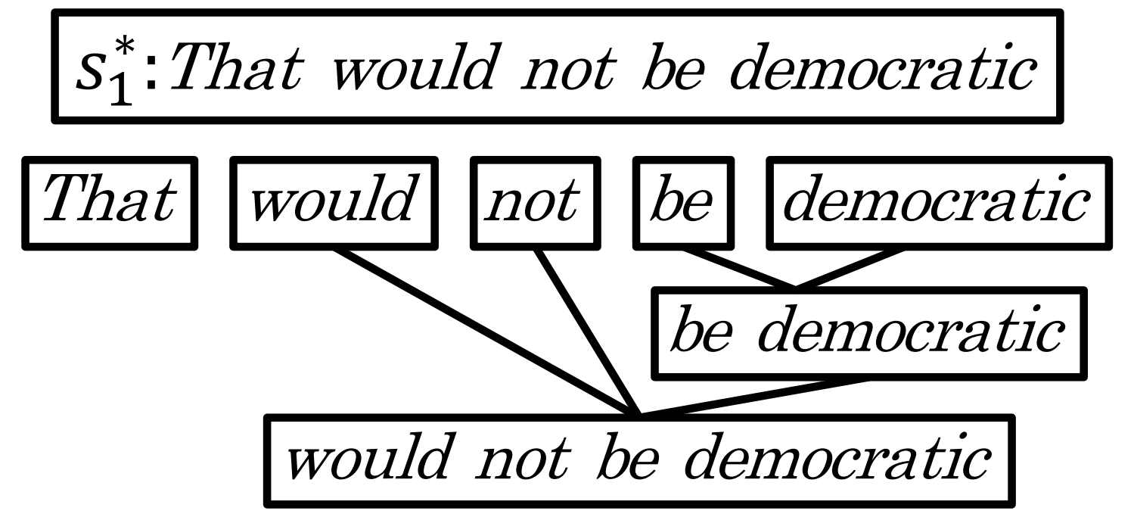


$\boldsymbol{s}^* = \operatorname{argmax}_{\boldsymbol{s}} \boldsymbol{a}_{\boldsymbol{u}}(\boldsymbol{s})$

$\boldsymbol{s}^*$ for Unit $\boldsymbol{u}$

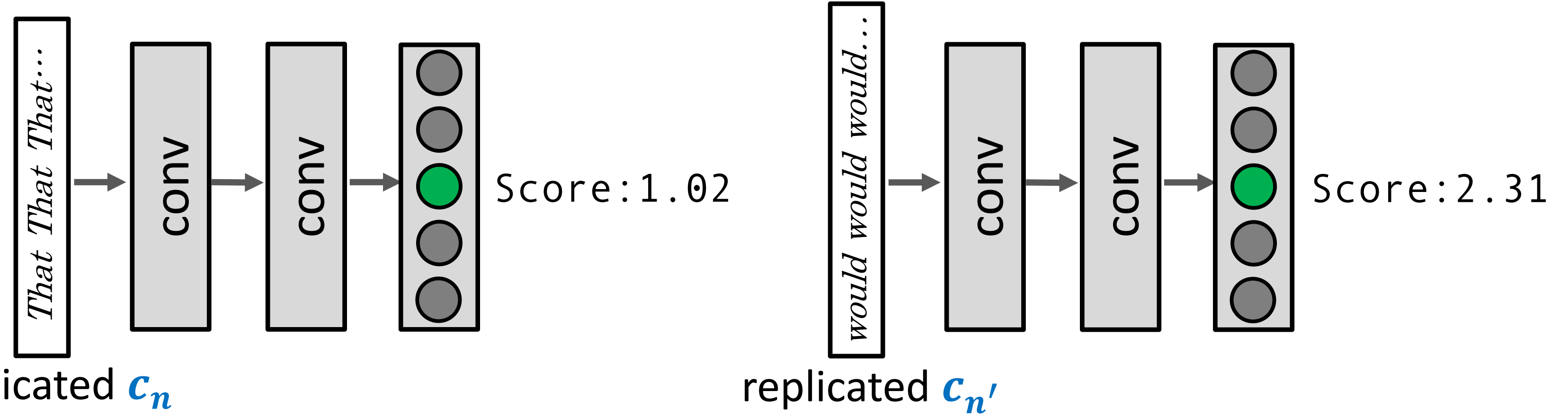
s_1^*: *That would not be democratic.*
s_2^*: *That would be cheap and it would not be right.*
s_3^*: *This is not how it should be in a democracy.*

3. Obtain candidate concepts from **constituency parse tree** of top k sentences $\boldsymbol{s}_{\boldsymbol{k}}^*$



Candidate Concepts
$C_u = \{c_1\text{: “That”}, c_2\text{: “would”}, \dots, c_N\text{: “woud not be democratic”}\}$

4. Compute `alignment_score(concept` $\boldsymbol{c_n}$`, unit` $\boldsymbol{u}$`) =` $a_{\boldsymbol{u}}$`(replicated` $\boldsymbol{c_n}$`)`



Which Concepts are Sensitive to Each Unit?

Layer14, Unit 690: what, who, where

- Who gets what, how much and when?
- On what basis, when and how?
- Then we need to ask: where do we start?
- However, what should we do at this point?
- What I am wondering now is: where are they?

Layer14, Unit 224: sure, know, aware

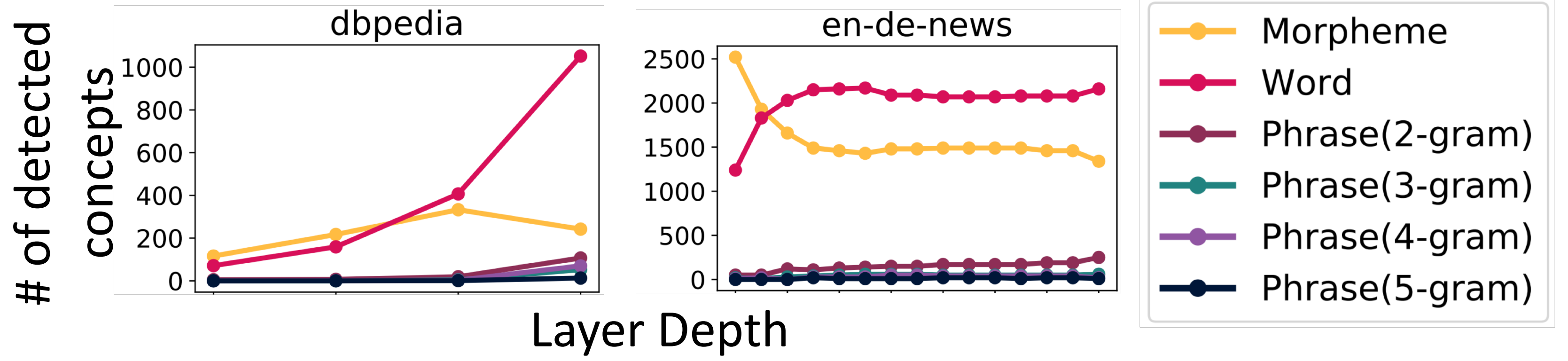
- Are you sure you are aware of our full potential?
- They know that and we know that.
- I am sure you will understand.
- I am sure you will do this.
- I am confident that we will find a solution.

Layer03, Unit 244: very disappointing, absolute worst place

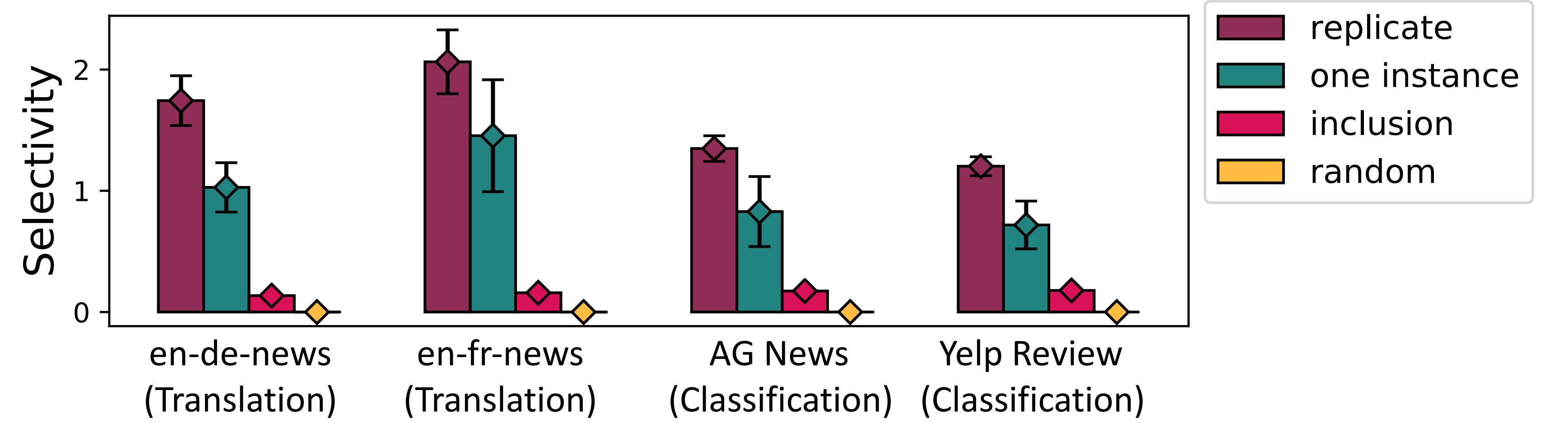
- very disappointing. ordered a vegetarian entrée,…
- what the hell did i pay for?…
- the absolute worst place i have ever done business with!
- the is by far the worst restaurant i have ever been to…
- this place is a rip off!…

- These units can serve as detectors for specific natural language concepts
- There are units capturing syntactically or semantically related concepts

Concept Granularity Evolves with Layer



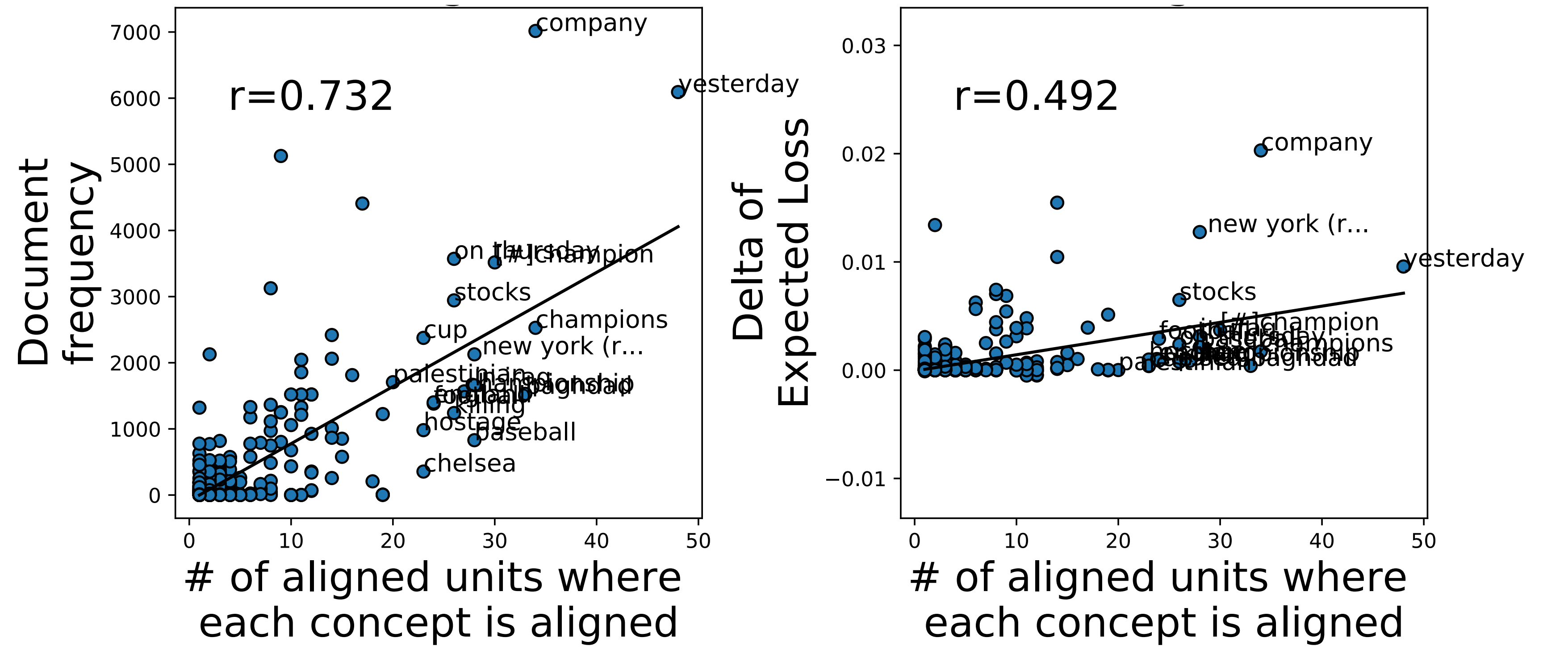
How Selectively does Each Unit Respond to Aligned Concepts?



$$\text{Selectivity}(\text{unit } \boldsymbol{u}) = \frac{\mu_+ - \mu_-}{\max_{s \in S} a_{\boldsymbol{u}}(s) - \min_{s \in S} a_{\boldsymbol{u}}(s)},$$
$$\text{where } \mu_+ = \frac{1}{|S_+|} \sum_{s \in S_+} a_{\boldsymbol{u}}(s), \quad \mu_- = \frac{1}{|S_-|} \sum_{s \in S_-} a_{\boldsymbol{u}}(s)$$

- Units are selectively responsive to specific concepts
- Our method successfully aligns such concepts to units

Which Concepts Appear More often?



Concepts that (1) **appear more often** in training data & (2) have **more influence on loss value** are detected in more units