Discovering Latent Concepts in BERT

Fahim Dalvi* Abdul Rafae Khan* Firoj Alam Nadir Durrani Jia Xu Hassan Sajjad

{faimaduddin,fialam,ndurrani,hsajjad}@hbku.edu.qa {akhan4,jxu70}@stevens.edu

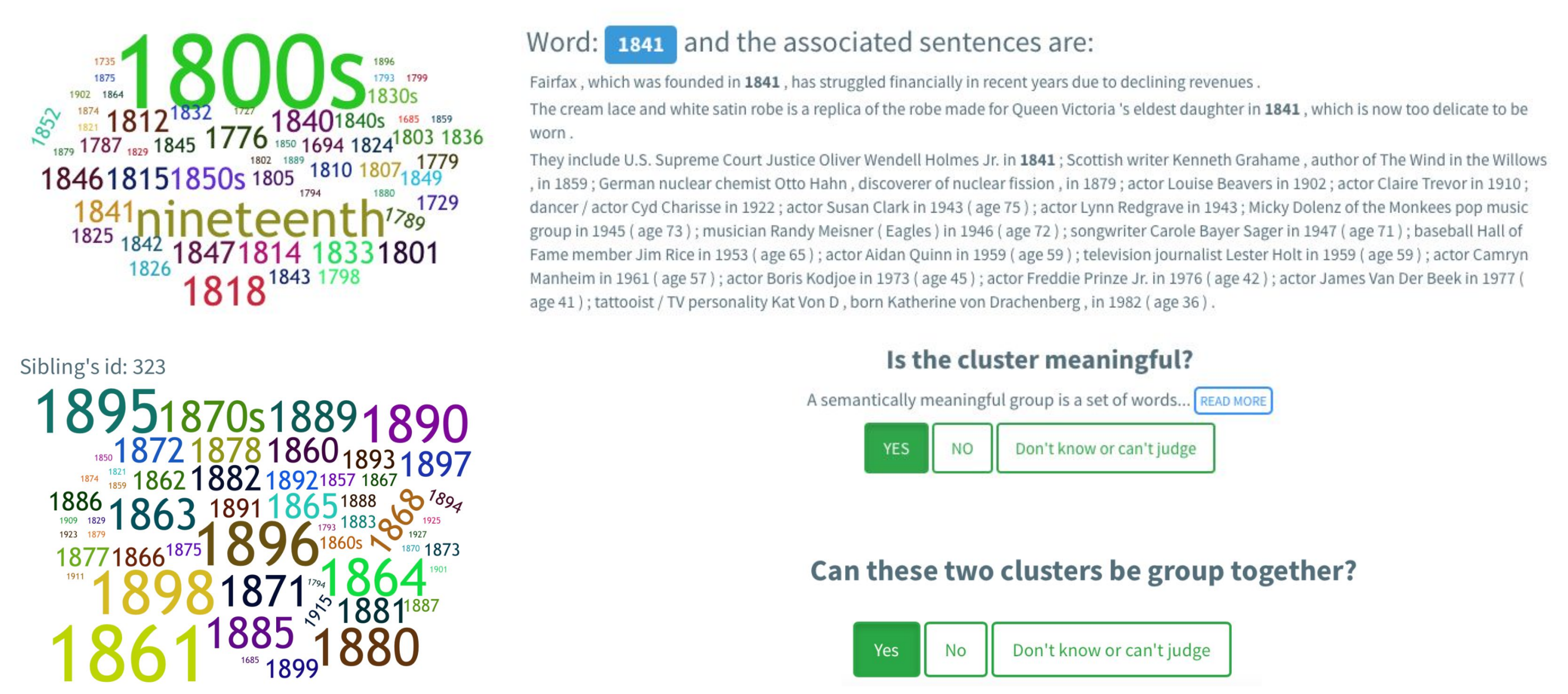
1. Introduction

- Current research on interpreting Deep NLP models is limited to pre-defined concepts
  - Classical NLP tasks (POS, NER, Chunking etc)
  - Ignores what latent concepts are learned by the model
- We propose a method to analyze latent concepts learned in pre-trained models
  - What are the novel concepts learned by the model?
  - How much do these latent concepts align with pre-defined linguistics concepts?
  - How do concepts evolve across the network layers
- We annotated latent concepts in BERT and provide a multi-facet hierarchical conceptNet dataset (BCN)
  - 174 fine-grained concepts and a total of 1M annotated instances
  - The dataset enables model-centric interpretation
  - The dataset can be used as a new classification dataset for NLP in general

2. Methodology

- **Concept**
  - represents a notion and can be viewed as a coherent fragment of knowledge
  - a group of words that are meaningful e.g. *Names of ice-hockey teams, First words of a sentence, Words that begin with "anti"*
- **Methodology**
  - Given a pre-trained model and a corpus of sentences
  - Extract contextualized representations of words
  - Group words into clusters using hierarchical clustering
  - Manually annotate each cluster into fine-grained categories
  - Analyze the cluster by aligning them with the pre-defined linguistic concepts

3. Annotation Task



Word: 1841 and the associated sentences are:

Fairfax , which was founded in **1841** , has struggled financially in recent years due to declining revenues .

The cream lace and white satin robe is a replica of the robe made for Queen Victoria 's eldest daughter in **1841** , which is now too delicate to be worn .

They include U.S. Supreme Court Justice Oliver Wendell Holmes Jr. in **1841** ; Scottish writer Kenneth Grahame , author of The Wind in the Willows , in 1859 ; German nuclear chemist Otto Hahn , discoverer of nuclear fission , in 1879 ; actor Louise Beavers in 1902 ; actor Claire Trevor in 1910 ; dancer / actor Cyd Charisse in 1922 ; actor Susan Clark in 1943 (age 75) ; actor Lynn Redgrave in 1943 ; Micky Dolenz of the Monkees pop music group in 1945 (age 73) ; musician Randy Meisner (Eagles) in 1946 (age 72) ; songwriter Carole Bayer Sager in 1947 (age 71) ; baseball Hall of Fame member Jim Rice in 1953 (age 65) ; actor Aidan Quinn in 1959 (age 59) ; television journalist Lester Holt in 1959 (age 59) ; actor Camryn Manheim in 1961 (age 57) ; actor Boris Kodjoe in 1973 (age 45) ; actor Freddie Prinze Jr. in 1976 (age 42) ; actor James Van Der Beek in 1977 (age 41) ; tattooist / TV personality Kat Von D , born Katherine von Drachenberg , in 1982 (age 36) .

Is the cluster meaningful?

A semantically meaningful group is a set of words... READ MORE

YES | NO | Don't know or can't judge

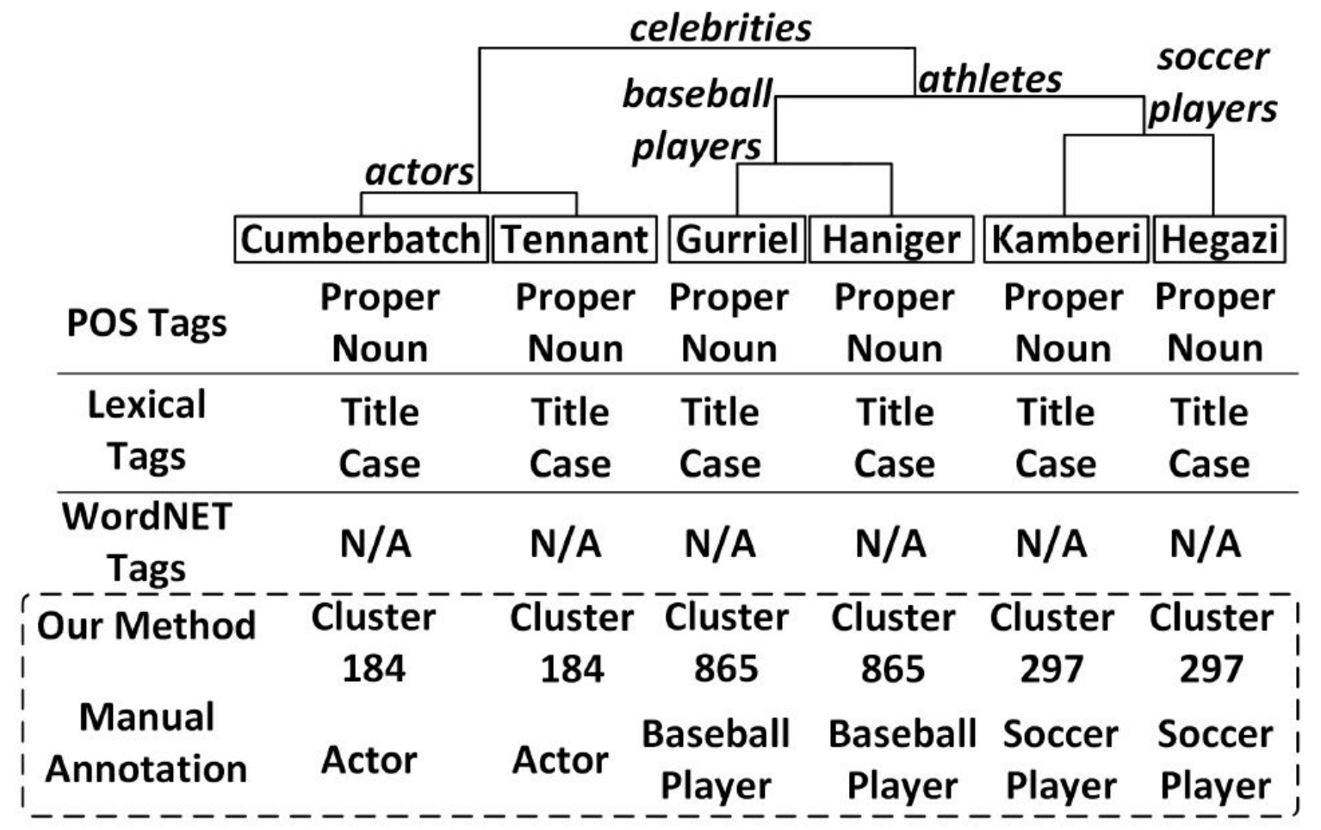
Can these two clusters be group together?

Yes | No | Don't know or can't judge

- We annotated 279 clusters
  - 243 (87.1%) meaningful and 36 (12.9%) non-meaningful clusters **(Q1)**
  - 142 (75.9%) can be combined with the sibling to form a bigger meaningful cluster **(Q2)**
- BCD dataset consists of 174 unique concepts
  - 11 lexical labels, 10 morphological labels, 152 semantic labels, and 1 syntactic label

5. Annotated Dataset

The annotation task preserves the concept hierarchy.

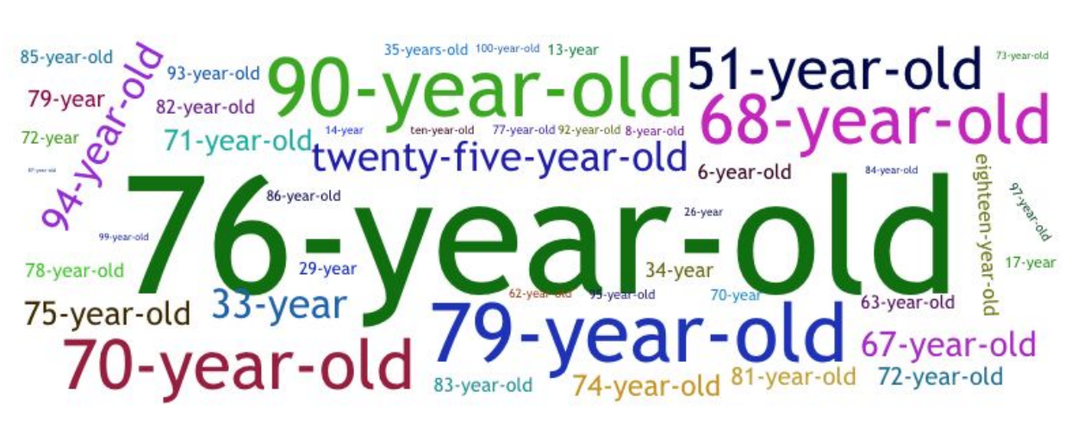


	Cumberbatch	Tennant	Gurriel	Haniger	Kamberi	Hegazi
POS Tags	Proper Noun	Proper Noun	Proper Noun	Proper Noun	Proper Noun	Proper Noun
Lexical Tags	Title Case	Title Case	Title Case	Title Case	Title Case	Title Case
WordNET Tags	N/A	N/A	N/A	N/A	N/A	N/A
Our Method	Cluster 184	Cluster 184	Cluster 865	Cluster 865	Cluster 297	Cluster 297
Manual Annotation	Actor	Actor	Baseball Player	Baseball Player	Soccer Player	Soccer Player

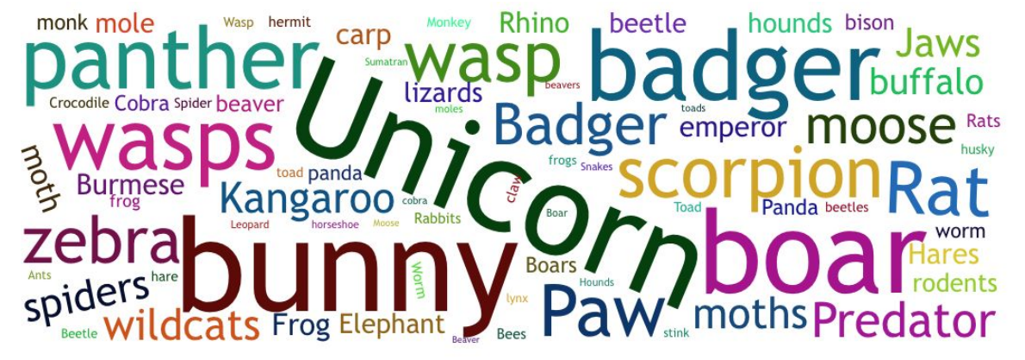
The multifaceted nature captures diverse information.

Brandenburg
Dresden
Stuttgart
Mainz
Trier
Weimar
Schleswig-Holstein
Heidelberg
Prussian
Innsbruck
Pomerania
Nuremberg
Freiburg
Augsburg
Bremen
Hanover

Semantic:named_entities:city
Semantic:geography:germany



Morphology:adjective
Lexical:hyphenation



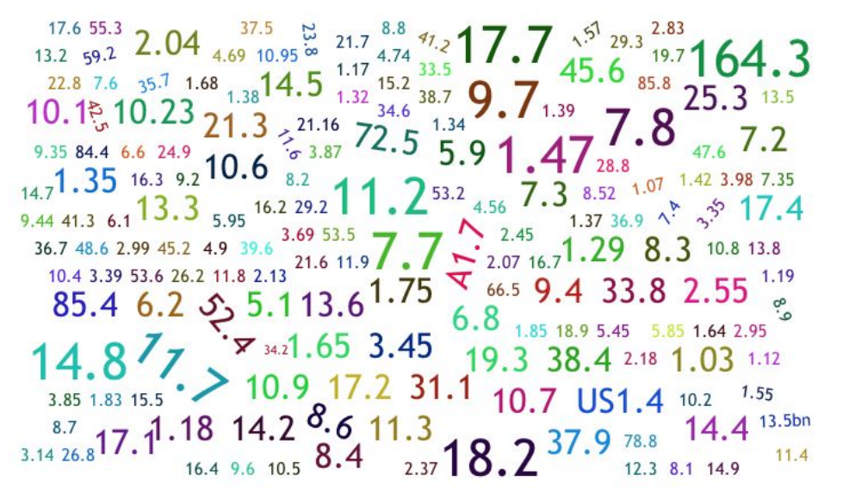
Semantic:animal
:land_animal
Morphology:noun

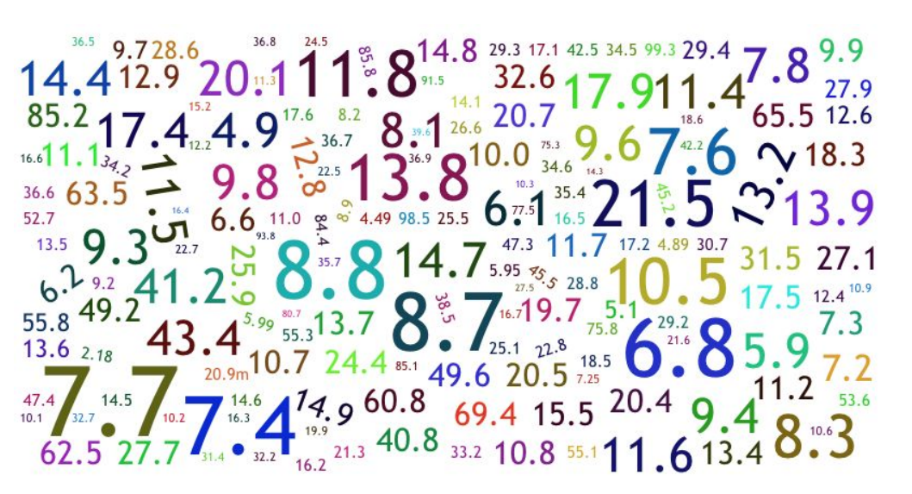
unplanned
undiscovered
unhelpful
unaffordable
unflinching
relatable
unambiguous
unmatched
infrequently
underrecognized
unregistered
unanimity
inscrutable
unrivaled
unelected
unverifiable

Lexical:prefix:un

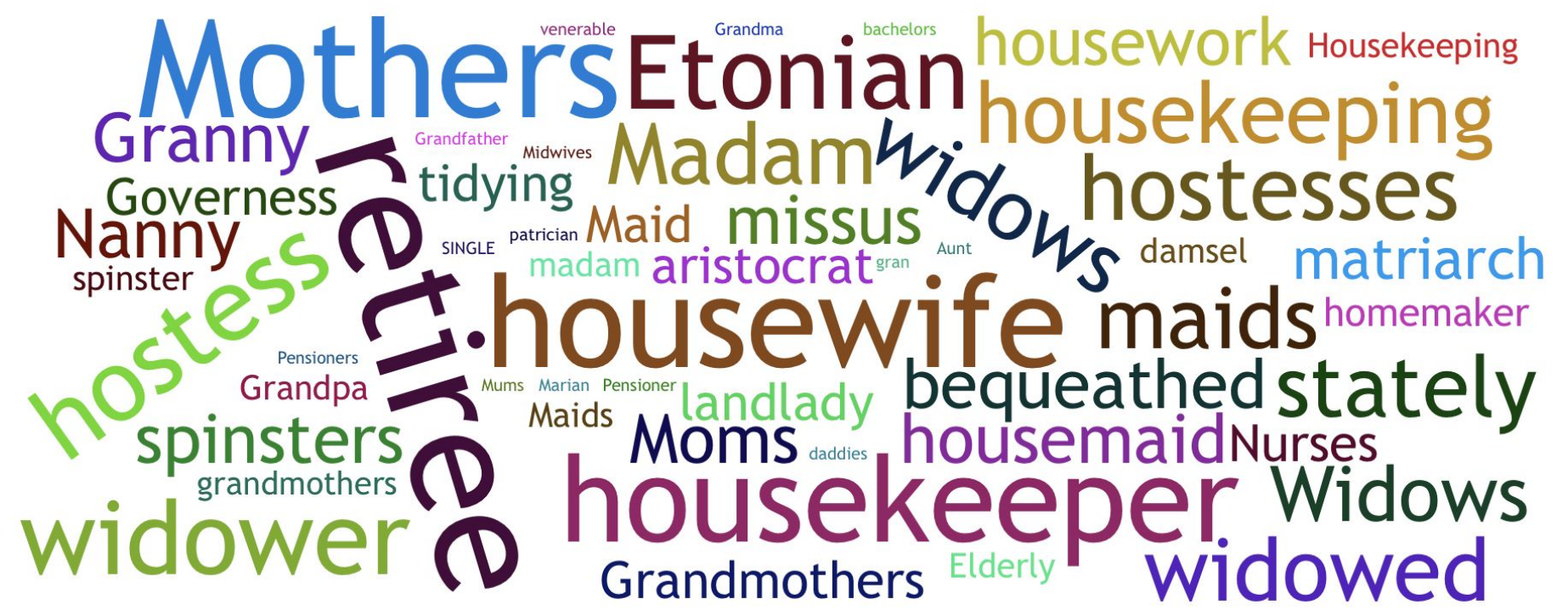
6. Analysis

- Lexically similar but semantically different clusters based on the context
  - Decimal numbers that capture monetary values, e.g €9.6, $2.4M
  - Decimal numbers which appear as percentages, e.g. 9.6%, 2.4%





- Cluster shows potential biases present in the training data
  - Female roles such as mother, aunt, granny are grouped together with specific job roles such as housekeeper, maid and nanny



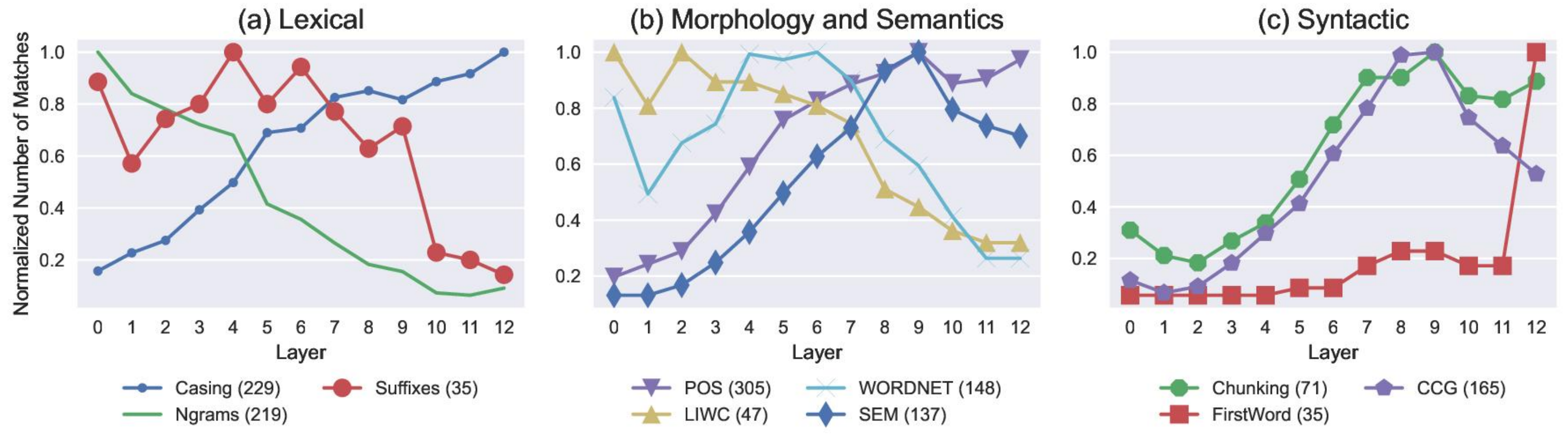
7. Alignment with pre-defined Concepts

- How much do latent BERT concepts align with pre-defined concepts ?
- Training data is annotated with pre-defined concepts
- A latent cluster is said to be aligned with the pre-defined concept if >=90% of its tokens belong to the pre-defined concept

	Lexical			Morphology and Semantics				Syntactic		
Concepts	Ngram	Suffix	Casing	POS	SEM	LIWC	WordNet	CCG	Chunk	FW
Matches	20 (2.0%)	5 (0.5%)	229 (23%)	297 (30%)	96 (10%)	15 (1.5%)	39 (3.9%)	87 (8.7%)	63 (6.3%)	35 (3.5%)

Alignment of BERT concepts for layer 12 with pre-defined concepts

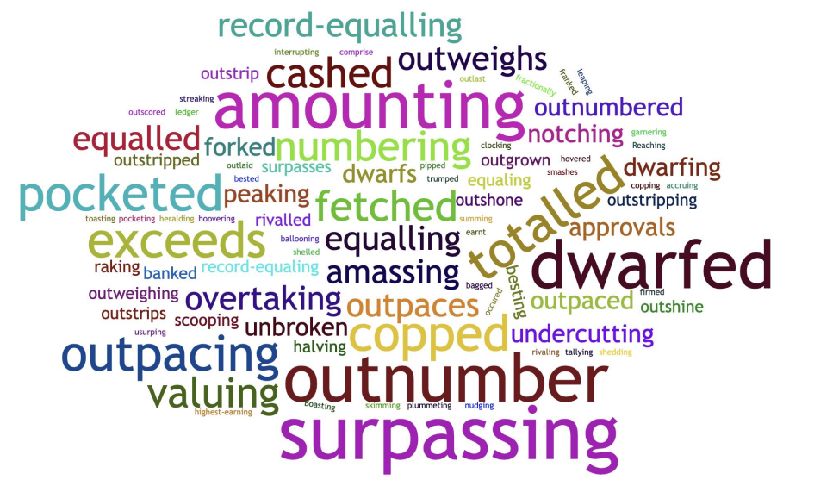
Evolution of Concepts Across Layers



- Lower layers encode the lexical and meaning-related knowledge
- The encoded concepts evolve into representing linguistic hierarchy, in the higher layers, taking contextual information into account
- Higher alignment with lexical concepts (e.g. suffixes) in the lower layers
- Higher alignment with psycholinguistic concepts (e.g. LIWC) in the initial and middle layers
- Classical NLP concepts (e.g. POS, SEM, Chunking are captured in the middle and final layers

8. Unaligned Clusters

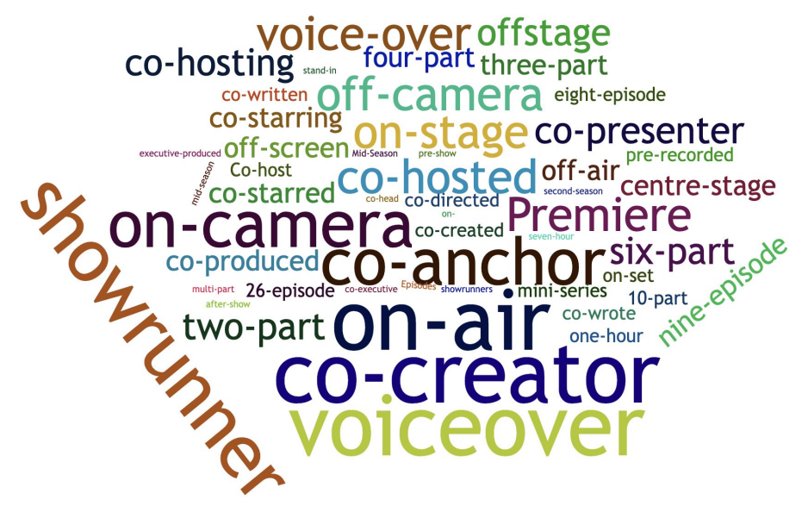
- What do the unaligned clusters represent?
- Compositionitonal clusters:
  - Figure (a) Verb forms and (b) Singular/Plural Nouns
- Unaligned but explainable: Figure (c) Compound Words
- Uninterpretable Clusters: Figure (d) No meaningful relation



(a) Verb Forms

herbal
cinnamon
herb
Coconut
almond
twigs
mango
pineapple
pea
indigo
Potato
blueberries
eucalyptus
hazelnut

(b) Singular and Plural Nouns



(c) Compound Words

cascading
thorny
churn
coax
snag
hiccups
niggling
wriggle
snaking
nagging
snowballed
squirm
hiccup
nab

(d) Uninterpretable

9. BCN Dataset

To expand the manually annotated data:

- We trained a logistic classifier on the annotated concepts
- Predict the cluster id of new tokens from a large News data
- We only select a prediction when the classifier is 97% confident about its prediction
- **BCN consists of 174 concept labels and a total of 1M annotated instances**

Algorithm 3 Concept Prediction with Logistic regression

Input: X_{train} = word representations for train data, Y_{train} = cluster id from Algorithm 1, X_{test} = word representations for test data,
Parameter: t = probability threshold

```
c = unique cluster ids from $Y_{train}$
M = train Logistic Regression model on $X_{train}$ and $Y_{train}$
for each $x \in X_{test}$ do
    $p$ = predict $K$ probabilities for each cluster using $M$ and input $x$
    $i = \arg\max p$
    if $p_i >= t$ then
        assign $x$ to cluster id $c_i$
    end if
end for
```

Application of BCN for Neuron Interpretation

- Discover neurons learning a pre-defined concept
- BCN provides fine-grained concepts e.g. person names are split into finer categories based on geography
- Select a fine-grained concept; muslim names and a coarse concept person names
- Identify the neurons responsible for each concept
- We used Linguistic Correlation Analysis from the NeuroX toolkit to identify minimum number of neurons required for the concept
- We found only 19 neurons were required for the concept muslim names compared to 74 required for the concept person
- Therefore, this shows that BCN helps enable selection of specialized neurons responsible for very specific aspects of language

Acknowledgements: This research was carried out in collaboration between the HBKU Qatar Computing Research Institute (QCRI) and the Stevens Institute of Technology (SIT) -- * denotes equal contribution