An Approximate Perspective on Word Prediction in Context: Ontological Semantics meets BERT

Kanishka Misra and Julia Taylor Rayz
Purdue University

NAFIPS 2020

Virtually, from West Lafayette, IN, USA

Summary and Takeaways

- Neural Networks based Natural Language Processing:

  Word Prediction in Context (WPC) -> Language Representations -> Tasks

- **This work:** Qualitative Account of WPC using a meaning-based approach to knowledge representation.
- Case Study on the BERT model (Devlin et al., 2019).

Word Prediction in Context

I unlocked the door using a ______.

- key
- lock-pick
- screwdriver
- ...

Cloze Tasks (Taylor, 1965)

Participants predict blank words in a sentence by relying on the context surrounding the blank.

Pretraining

Process of training a Neural Network on large texts. Usually using a Language Modelling objective

word

Trainable parameters

For a sequence of length T: $$\sum_{t=0}^{T} \log P(w_t \mid h_{t-1}, \Theta_{t-1})$$

Hidden state (representations useful for NL tasks)

Misra and Rayz, 2020

BERT - Bidirectional Encoder Representations from Transformers

Large Transformer network (Vaswani et al., 2017) trained on large pieces of text to do the following:

Oh, I love coffee! I take coffee with **[MASK]** and sugar.
1 2

1) **Masked Language Modelling:** What is [MASK]?
2) **Next Sentence Prediction:** Does **2** follow **1**?

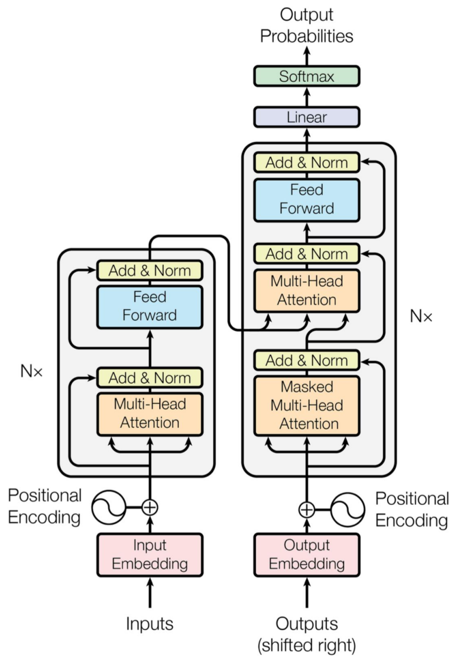


Figure 1: The Transformer - model architecture.

(Figure from Vaswani et al., 2017)

Devlin et al., 2019

Misra and Rayz, 2020

Semantic Capacities of BERT

Strong empirical performance when tested on:

- **Attributing nouns to their hypernyms:** *A robin is a bird.*
- **Commonsense and Pragmatic Inference:** *He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of [MASK].*

P(football) > P(chess)

- **Lexical Priming:**
  - (1) ***delicate.*** *The tea set is very [MASK].*
  - (2) ***salad.*** *The tea set is very [MASK].*

P(**fragile** | (1)) > P(**fragile** | (2))

(Ettinger, 2020; Petroni et al., 2019; Misra et al., 2020)

Misra and Rayz, 2020

Semantic Capacities of BERT

Weak performance when tested on:

- **Role-reversal:** *waitress serving customer* vs. *customer serving waitress.*
- **Negation:** *A robin is not a [MASK]*. P(bird) = high.

To what extent does BERT understand Natural Language?

(Ettinger, 2020; Kassner and Shutze, 2020)

Misra and Rayz, 2020

Analyzing BERT's Semantic and World Knowledge Capacities

Commonsense & World Knowledge

Items adapted from Psycholinguistic experiments (Ettinger, 2020):

Federmeier and Kutas (1999): *He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of [MASK].*

P(football|context) > P(chess|context) **[~75% accuracy]**

Items constructed from existing Knowledge bases (Petroni et al., 2019)

iPod Touch was produced by [MASK].

Argmax P([MASK] = x) = Apple

Misra and Rayz, 2020

Analyzing BERT's Semantic and World Knowledge Capacities

Semantic Inference

Items adapted from Psycholinguistic experiments (Ettinger, 2020):

Chow et al. (2016): (1) *the restaurant owner forgot which customer the waitress had [MASK].*
(2) *the restaurant owner forgot which waitress the customer had [MASK].*

P([MASK] = served | (1)) > P([MASK] = served | (2)) **[~80% accuracy]**

Fischler et al. (1983): (1) *A robin is a [MASK].*
(2) *A robin is not a [MASK].*

<add results>

Misra and Rayz, 2020

Analyzing BERT's Semantic and World Knowledge Capacities

Lexical Priming

Items adapted from Semantic Priming experiments (Misra, Ettinger, & Rayz, 2020):

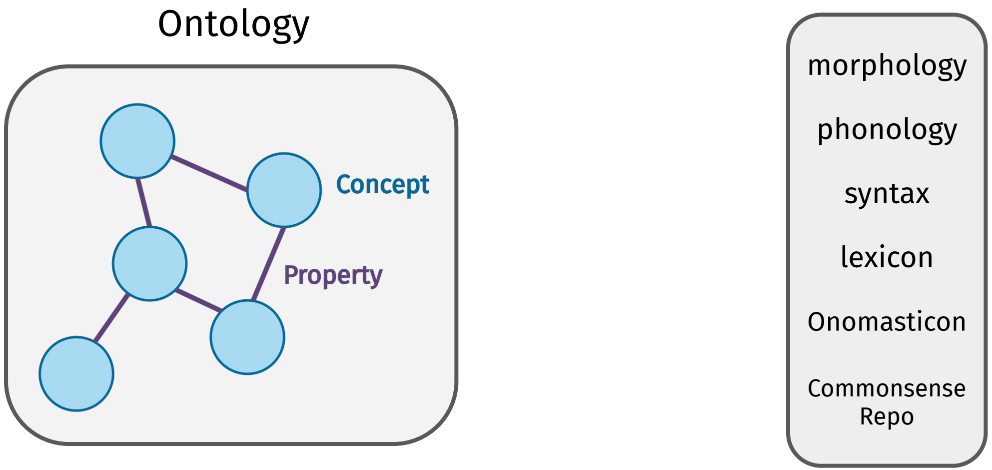
(1) ***delicate.*** The tea set was very [MASK].
(2) ***salad.*** The tea set was very [MASK].

<add results>

Misra and Rayz, 2020

Ontological Semantic Technology (OST)

Meaning first approach to knowledge representation (Nirenburg and Raskin, 2004).



Taylor, Raskin, Hempelmann (2010); Hempelmann, Raskin, Taylor (2010); Raskin, Hempelmann, Taylor (2010)

Fuzziness in OST

Facets assigned to properties of Events.

For any event, E, its facets represent memberships of concepts based on the properties that are endowed to E.

INGEST-1
- AGENT:
  - sem: ANIMAL
  - relaxable-to: SOCIAL-OBJECT
- THEME:
  - sem: FOOD, BEVERAGE
  - relaxable-to: ANIMAL, PLANT
  - not: HUMAN

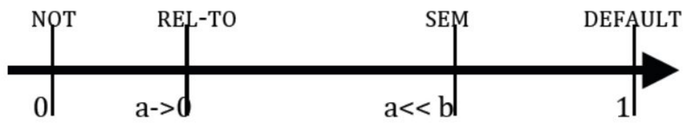


Fig. 2: Relative membership values of facet gradation

Taylor and Raskin (2010, 2011, 2016)

Misra and Rayz, 2020

Fuzziness in OST

Descendents of the default concept have higher membership than the sem facet.

E.g. TEACHER and INEXPERIENCED-TEACHER

$$\mu_E(\text{sem}) < \mu_E(descendant(\text{default})) < 1$$

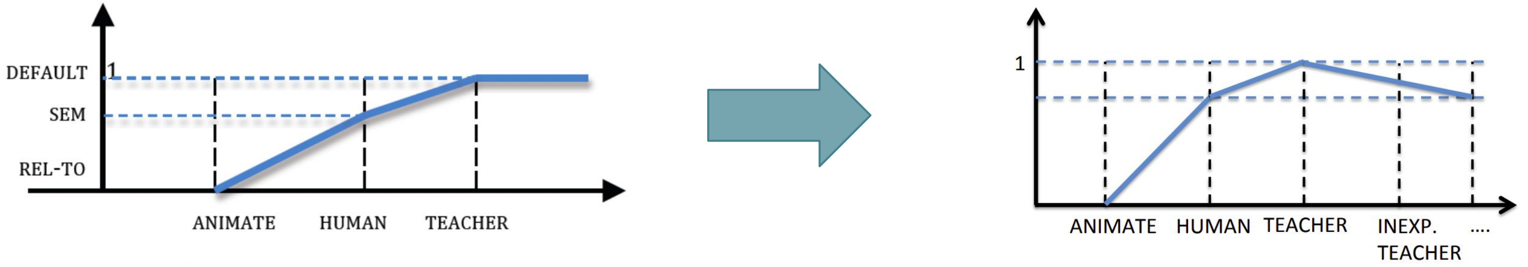


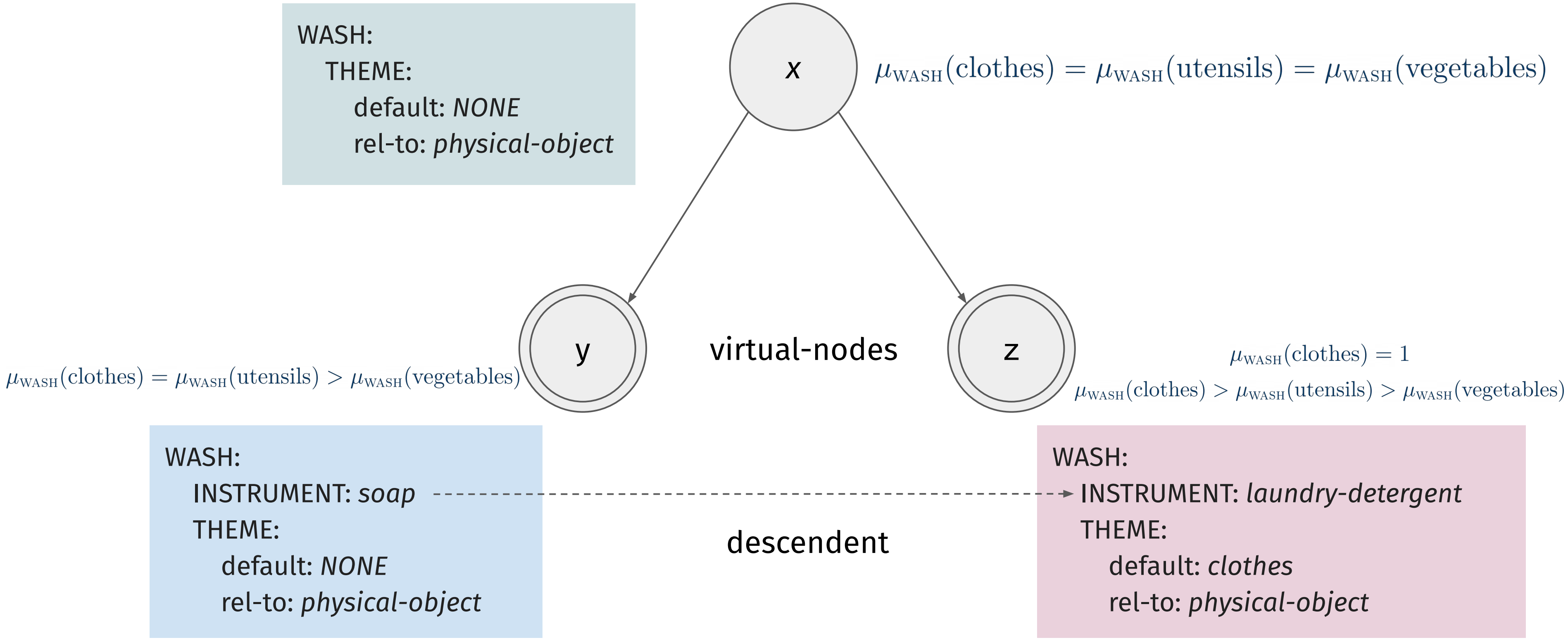
Fig. 4: (a) Fillers of agent TEACH with corresponding membership [10]

Fig. 4: (b) Fillers of agent TEACH with corresponding new membership

Calculation of μ : Taylor and Raskin (2010, 2011, 2016); Taylor, Raskin and Hempelmann (2011)

Misra and Rayz, 2020

Fuzziness in OST



Misra and Rayz, 2020

WPC as *Guessing the Meaning of an Unknown Word*

Using cloze tasks as the basis of learning the meaning of words is not new.

Taylor, Raskin, and Hempelmann (2010, 2011): OST and Cloze-tasks to infer the meaning of an unknown word.

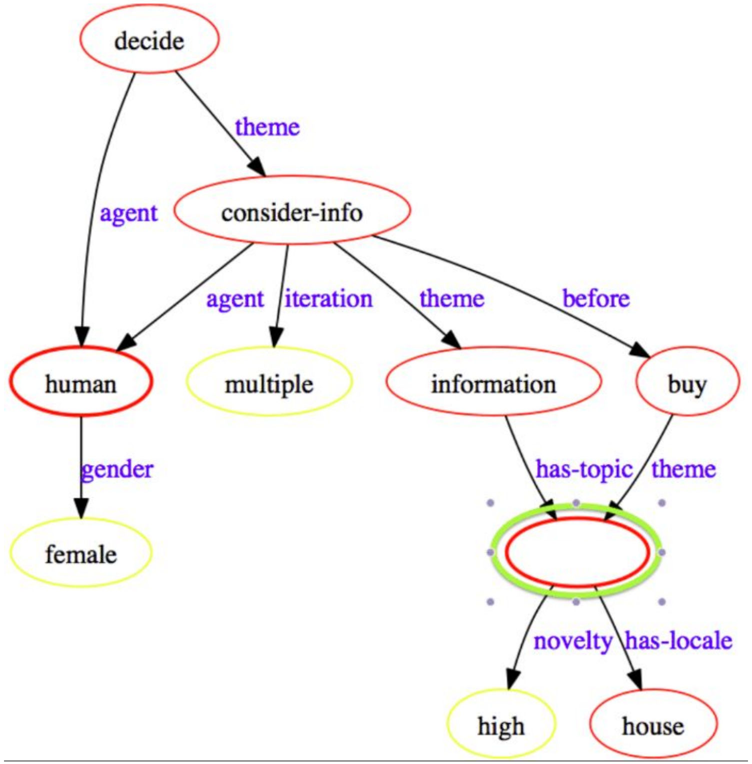
She decided she would rethink ***zzz*** before buying them for the whole house.

(the new curtains)

Misra and Rayz, 2020

WPC as *Guessing the Meaning of an Unknown Word*

She decided she would rethink ***zzz*** before buying them for the whole house.



Misra and Rayz, 2020

What is *zzz* according to BERT?

She decided she would rethink *zzz* before buying them for the whole house.

Rank	Token	Probability	Rank	Token	Probability
1	clothes	0.1630	21	design	0.0067
2	designs	0.1320	22	curtains	0.0063
13	paintings	0.0131	23	gifts	0.0060
16	furniture	0.0111	24	wardrobe	0.0057
17	pictures	0.0101	25	products	0.0049
18	books	0.0096	26	toys	0.0047
19	decorations	0.0078	28	photos	0.0041
20	arrangements	0.0070	30	decor	0.0040

Misra and Rayz, 2020

Interpreting an Example Sentence

She quickly got dressed and brushed her [MASK].

BRUSH:
AGENT: *HUMAN*
GENDER: *FEMALE*
THEME: [MASK]
INSTRUMENT: *NONE*

1. Act of cleaning [*brush your teeth*]
2. Rub with brush [*I brushed my clothes*]
3. Remove with brush [*brush dirt off the jacket*]
4. Touch something lightly [*her cheeks brushed against the wind*]
5. ...

Misra and Rayz, 2020

Interpreting an Example Sentence - BERT output

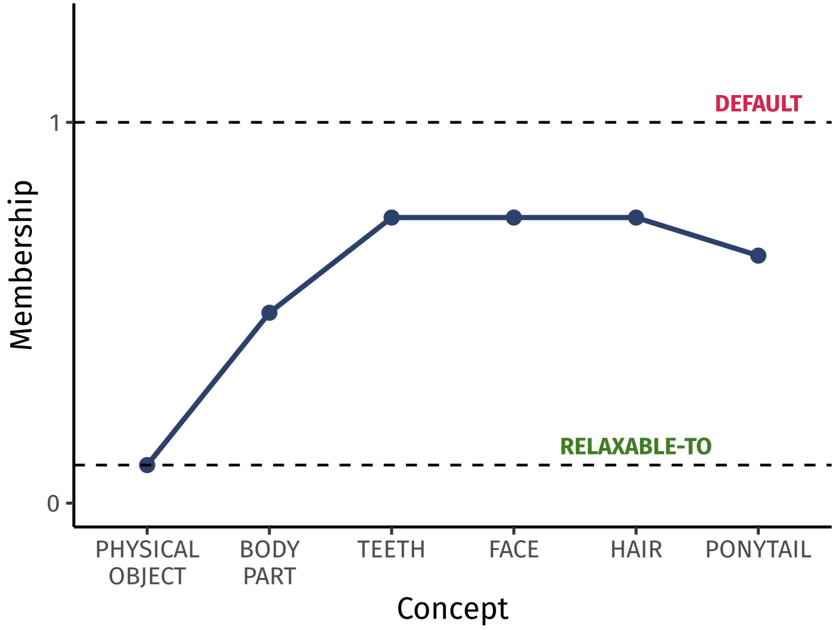
She quickly got dressed and brushed her [MASK].

Rank	Token	Probability
1	teeth	0.8915
2	hair	0.1073
3	face	0.0002
4	ponytail	0.0002
5	dress	0.0001

Misra and Rayz, 2020

Interpreting an Example Sentence - Emergent μ's

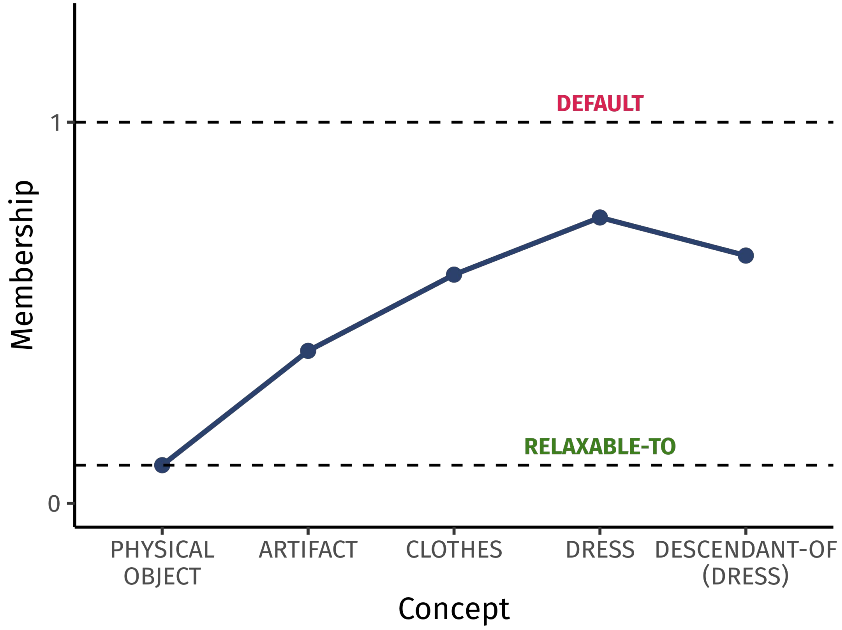
BRUSH-V1 with BODY-PART concepts as predicted completions



Misra and Rayz, 2020

Interpreting an Example Sentence - Emergent μ's

BRUSH-V1 with ARTIFACT concepts as predicted completions



Misra and Rayz, 2020

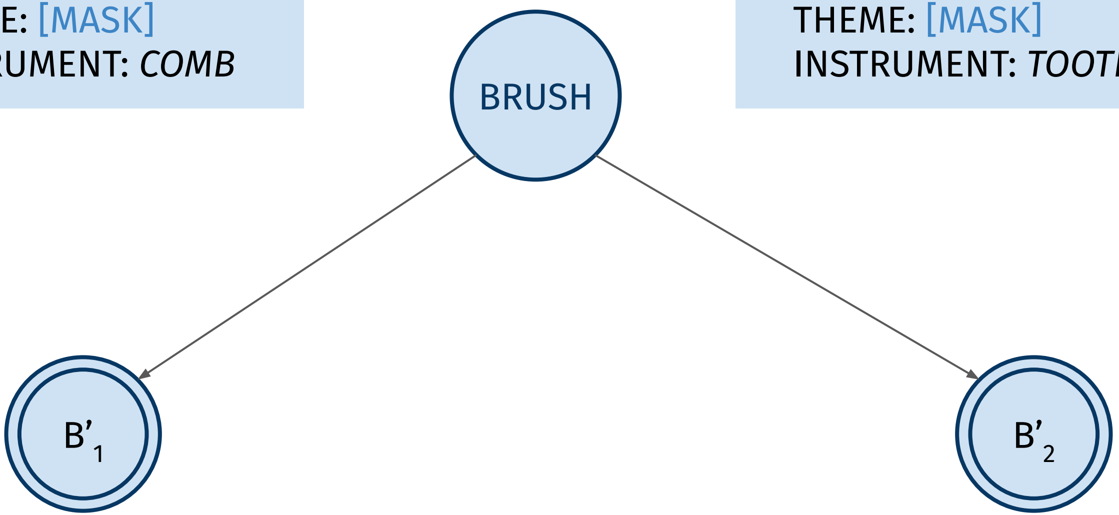
Interpreting an Example Sentence - More Properties!

She quickly got dressed and brushed her [MASK] with a comb.

She quickly got dressed and brushed her [MASK] with a toothbrush.

BRUSH:
AGENT: *HUMAN*
GENDER: *FEMALE*
THEME: [MASK]
INSTRUMENT: *COMB*

BRUSH:
AGENT: *HUMAN*
GENDER: *FEMALE*
THEME: [MASK]
INSTRUMENT: *TOOTHBRUSH*

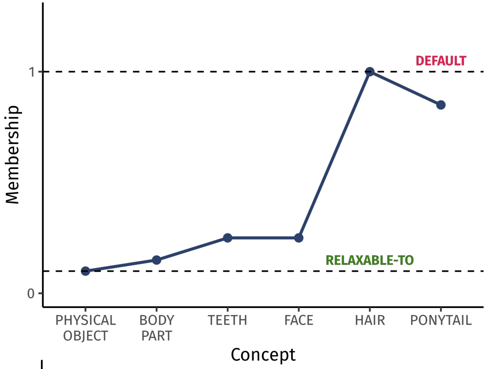


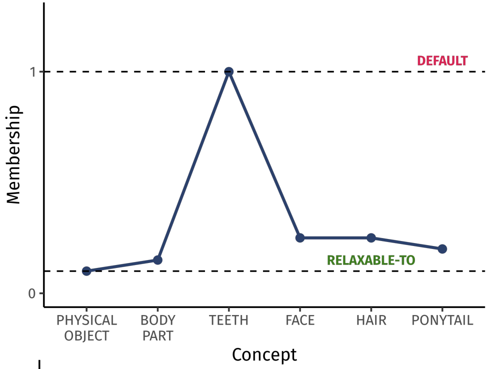
Interpreting an Example Sentence - More Properties!

She quickly got dressed and brushed her [MASK] with a comb.

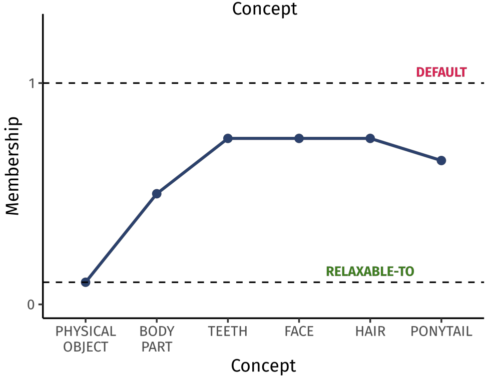
She quickly got dressed and brushed her [MASK] with a toothbrush.

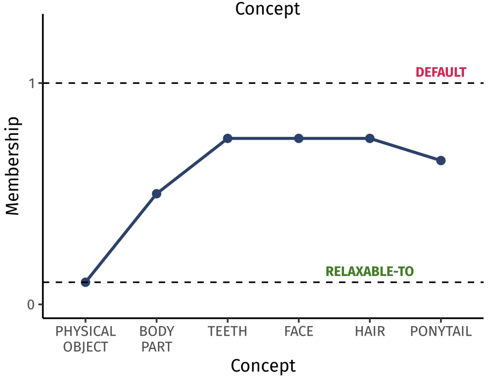
BRUSH-WITH-INSTRUMENT





BRUSH

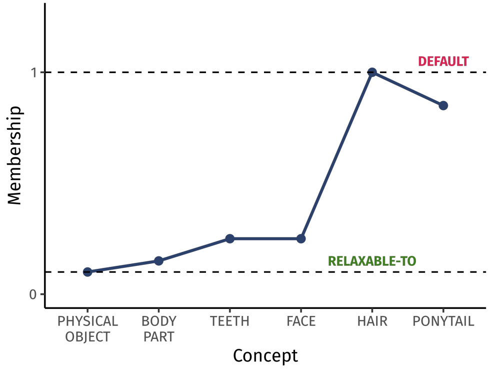




Interpreting an Example Sentence - More Properties!

She quickly got dressed and brushed her [MASK] with a comb.

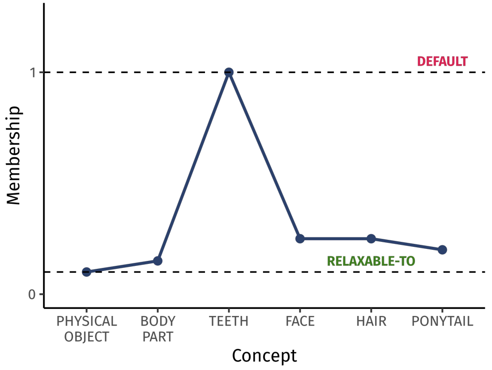
Rank	Token	Probability
1	hair	0.8704
2	teeth	0.1059
3	face	0.0210
12	ponytail	<0.0001
27	dress	<0.0001



BRUSH-WITH-INSTRUMENT

She quickly got dressed and brushed her [MASK] with a toothbrush.

Rank	Token	Probability
1	teeth	0.9922
2	hair	0.0052
3	face	0.0019
31	ponytail	<0.0001
98	dress	<<0.0001



Summary of Analysis

- BERT changes its top-predicted word when the instrument of the event changes.
- It is unable to show structural (semantics-wise) phenomena.
- **Evidence:** scoring descendent of HAIR, PONYTAIL lower than a nonsensical concept (in the given instance) – TEETH

Misra and Rayz, 2020

Summary and Takeaways

- BERT might be good at predicting defaults.
  - needs large scale empirical testing by collecting events and their defaults.
- BERT's MLM training procedure prevents it from learning equally plausible candidates of event fillers.
  - **Hypothesis:** Softmax isn't set up to learn multiple-labels per sample.
  - Especially when limited instances of the same event are encountered in training.
- Ontological Semantics provide semantic desiderata for word prediction in context using fuzzy inferences.

Misra and Rayz, 2020

Questions?