How do we get to general purpose NLU?

Emily M. Bender
University of Washington
@emilymbender

Google, Mountain View
5 September 2019

Acknowledgments

- This talk is based on joint work with Alexander Koller; Dan Flickinger, Stephan Oepen, Woodley Packard and Ann Copestake; and Alex Lascarides

- Photo (kittens & puppy): Image by JacLou DL from Pixabay

This talk in a nutshell

- A whole pile of end-to-end systems does not general-purpose NLU make
- Systems, no matter how complex, trained only on form, won't learn meaning
    - That's not how babies do it either
- General-purpose NLU requires attention to linguistic structure and use
- Compositionality is key!

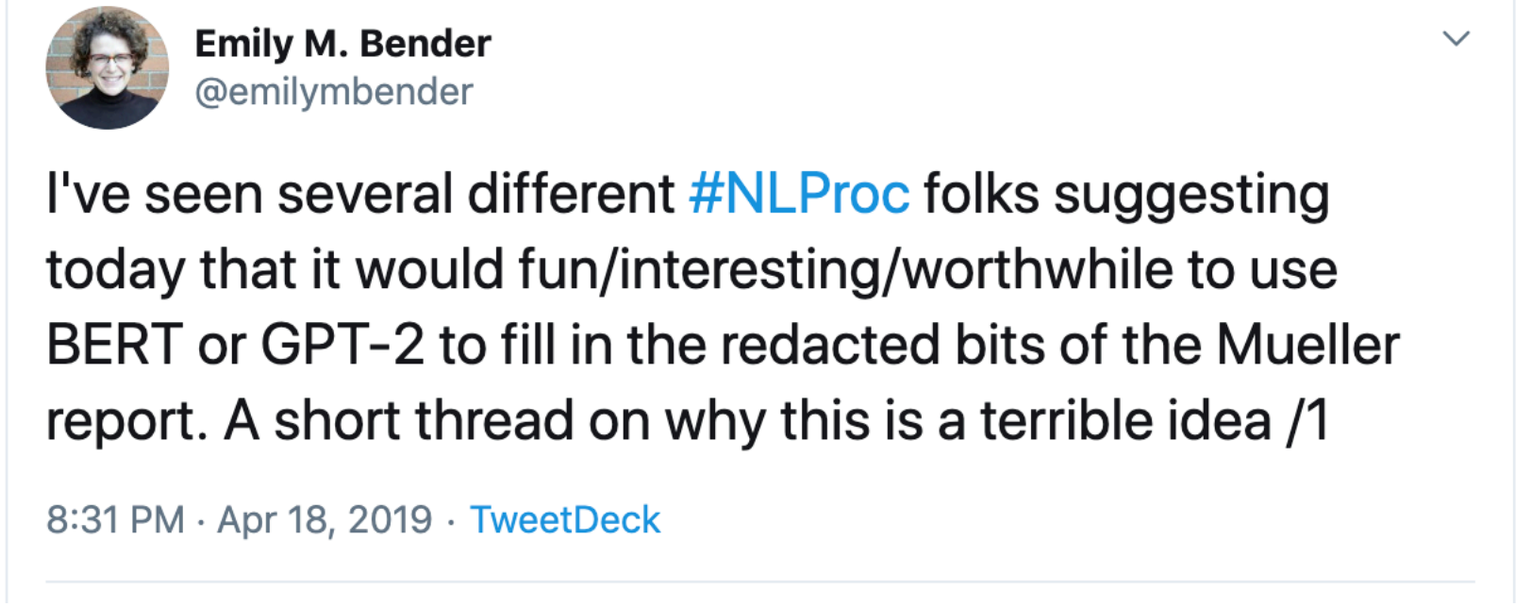
Outline

- Why this argument needs to be made
- Form v. meaning v. use v. world
- Linguistic knowledge (for developers and machines)
- Leveraging compositionality

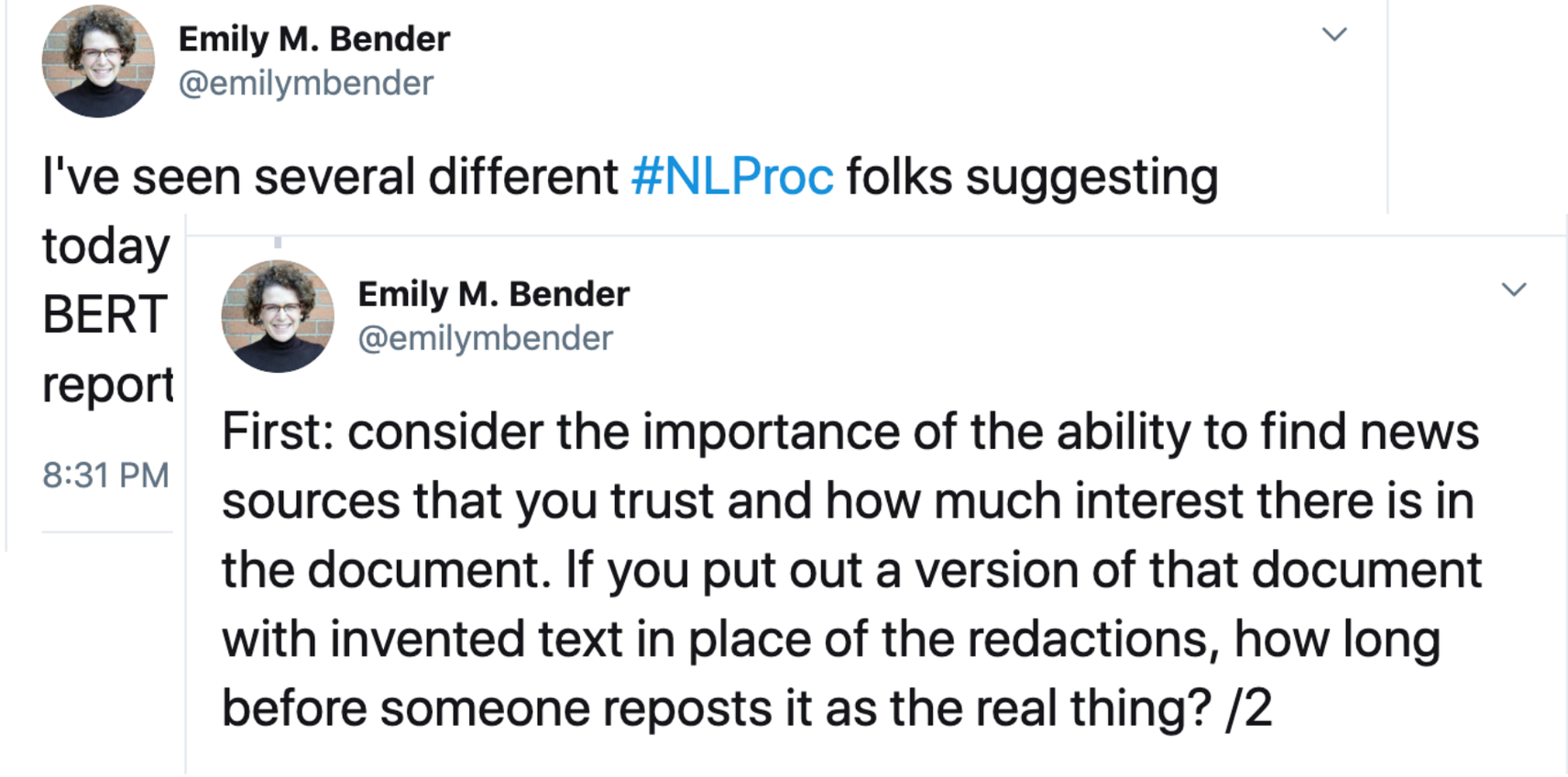
Outline

- **Why this argument needs to be made**
- Form v. meaning v. use v. world
- Linguistic knowledge (for developers and machines)
- Leveraging compositionality

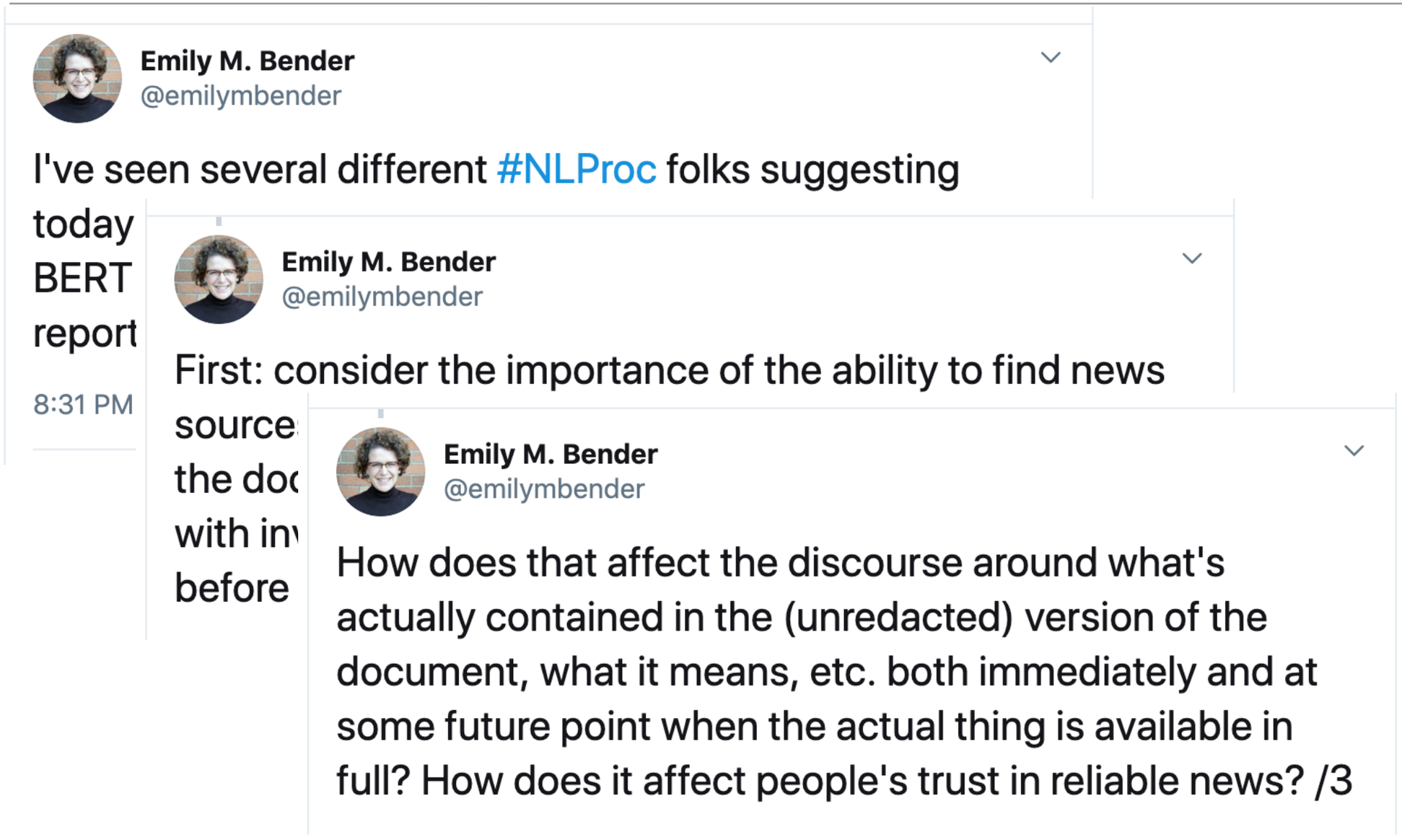
“Unredacting”



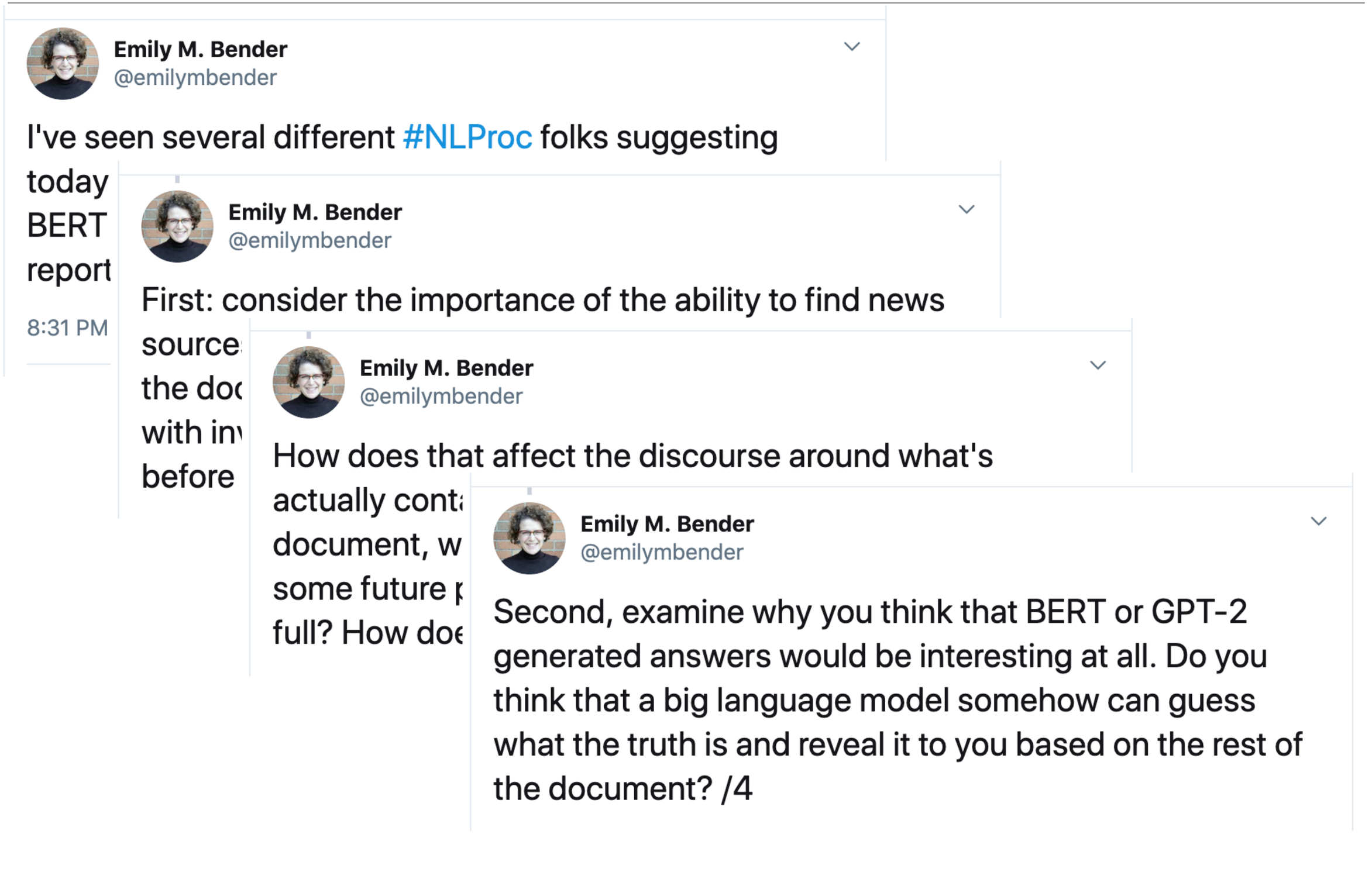
“Unredacting”



“Unredacting”



“Unredacting”



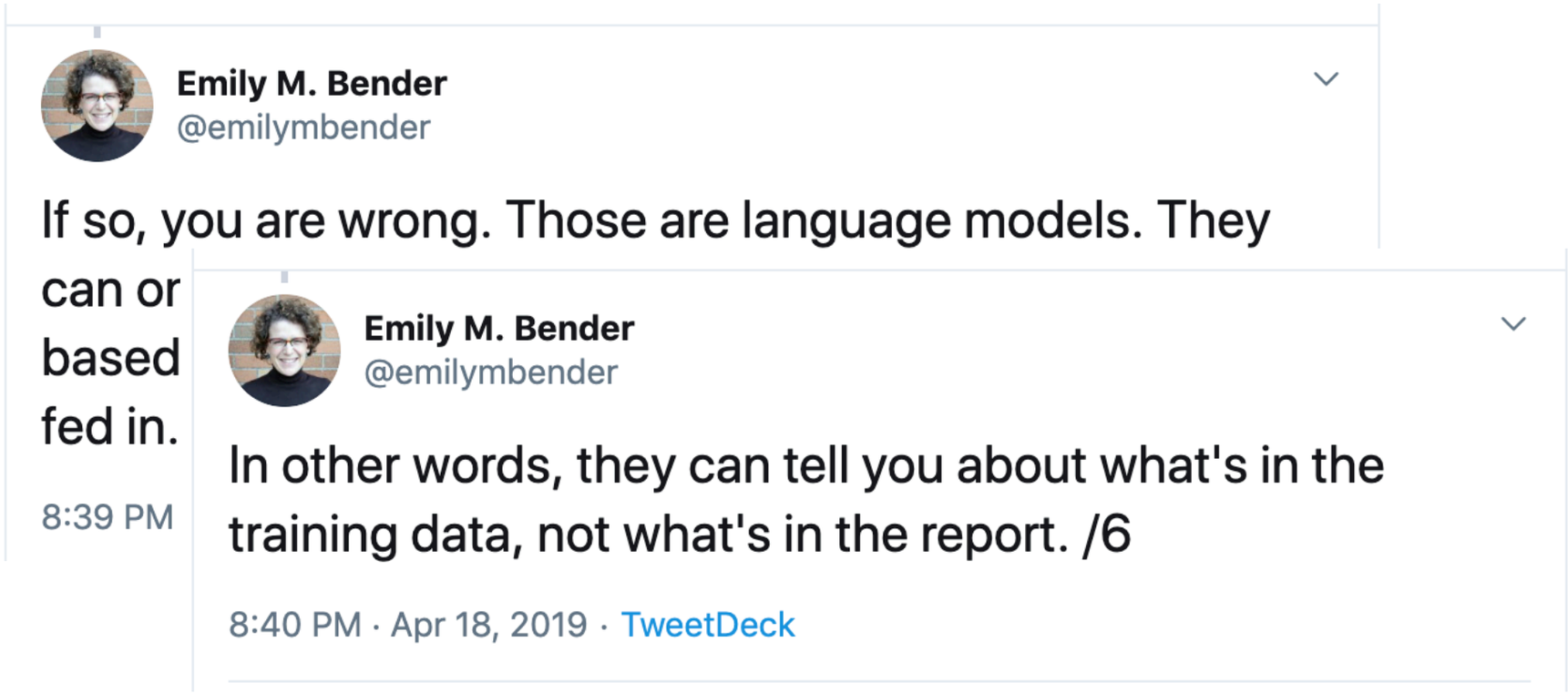
“Unredacting”

Emily M. Bender
@emilymbender

If so, you are wrong. Those are language models. They can only come up with sequences that are probable based on what's seen in the training data, given the prefix fed in. /5

8:39 PM · Apr 18, 2019 · TweetDeck

“Unredacting”



“Unredacting”

Emily M. Bender
@emilymbender

If so, you are wrong. Those are language models. They can or based fed in.

8:39 PM

Emily M. Bender
@emilymbender

In other words, they can tell you about what's in the trainir

8:40 PM

Emily M. Bender
@emilymbender

I haven't looked at the report, but I'm fairly confident the people doing the redacting would have been careful to do it in such a way that the redacted info furthermore is not predictable.

(And, ahem, that the black-out can't just be deleted...) /7

8:42 PM · Apr 18, 2019 · TweetDeck

“Unredacting”

Emily M. Bender
@emilymbender

If so, you are wrong. Those are language models. They can or based fed in.

8:39 PM

Emily M. Bender
@emilymbender

In other words, they can tell you about what's in the trainir

8:40 PM

Emily M. Bender
@emilymbender

I haven't looked at the report, but I'm fairly confident the people doi it in such a predictabl

(And, aher

8:42 PM · Apr

Emily M. Bender
@emilymbender

So, please, just stop it with this idea. It's not funny nor helpful. If you're interested in applying #NLProc in ways relevant to the current political moment, how about working on e.g. rumor detection and tools that might help users think twice before retweeting/sharing? /fin

8:44 PM · Apr 18, 2019 · TweetDeck

“NLP’s ImageNet moment has arrived” (Sebastian Ruder: https://thegradient.pub/nlp-imagenet/)

Word2vec and related methods are *shallow* approaches that trade expressivity for efficiency. Using word embeddings is like initializing a computer vision model with pretrained representations that only encode edges: they will be helpful for many tasks, but they fail to capture higher-level information that might be even more useful. A model initialized with word embeddings needs to learn from scratch not only to disambiguate words, but also to derive meaning from a sequence of words. This is the core aspect of language understanding, and it requires modeling complex language phenomena such as compositionality, polysemy, anaphora, long-term dependencies, agreement, negation, and many more. It should thus come as no surprise that NLP models initialized with these shallow representations still require a huge number of examples to achieve good performance.

“NLP’s ImageNet moment has arrived” (Sebastian Ruder: https://thegradient.pub/nlp-imagenet/)

Word2vec and related methods are *shallow* approaches that trade expressivity for efficiency. Using word embeddings is like initializing a computer vision model with pretrained representations that only encode edges: they will be helpful for many tasks, but they fail to capture higher-level information that might be even more useful. A model initialized with word embeddings needs to learn from scratch not only to disambiguate words, but also to derive meaning from a sequence of words. This is the core aspect of language understanding, and it requires mod
long-term de
surprise that
number of ex

In order to predict the most probable next word in a sentence, a model is required not only to be able to express syntax (the grammatical form of the predicted word must match its modifier or verb) but also model semantics. Even more, the most accurate models must incorporate what could be considered *world knowledge* or *common sense*. Consider the incomplete sentence "The service was poor, but the food was". In order to predict the succeeding word such as “yummy” or “delicious”, the model must not only memorize what attributes are used to describe food, but also be able to identify that the conjunction “but” introduces a contrast, so that the new attribute has the opposing sentiment of “poor”.

Child language development requires more than just exposure to language (with or without vision)

- Learning from text only is not “just like babies do it”
- Early language acquisition is predicated on *joint attention* (Bruner 1985, Tomasello & Farrar 1986, inter alia)
- Even phonetic learning requires social engagement, exposure via TV or radio alone is insufficient (Kuhl 2007)

Outline

- Why this argument needs to be made
- **Form v. meaning v. use v. world**
- Linguistic knowledge (for developers and machines)
- Leveraging compositionality

Form v. meaning v. use v. world

- Form: text, speech, sign (+ paralinguistic information like gesture or tone)

- Conventional/standing meaning: logical form (or equivalent) that the linguistic system pairs with that form

- Communicative intent of the speaker: what they are publicly committed to by uttering that form (+ additional plausibly deniable inferences)

- Relationship between communicative intent & the world, e.g.:

    - True assertion, mistaken assertion, lie, accidentally true assertion, social act related to construction of social world, question about the interlocutor's beliefs, …

Form v. meaning v. use v. world

Form v. meaning v. use v. world

- Form: ഒരു ഭാഷ ഒരിക്കലും മതിയാവില്ല

Form v. meaning v. use v. world

- Form: ഒരു ഭാഷ ഒരിക്കലും മതിയാവില്ല

- Conventional meaning: “One language is never enough”

Form v. meaning v. use v. world

- Form: ഒരു ഭാഷ ഒരിക്കലും മതിയാവില്ല
- Conventional meaning: “One language is never enough”
- Speaker intent: at a tourist market, after a long trip, while maintaining legacy code, …

Form v. meaning v. use v. world

- Form: ഒരു ഭാഷ ഒരിക്കലും മതിയാവില്ല
- Conventional meaning: “One language is never enough”
- Speaker intent: at a tourist market, after a long trip, while maintaining legacy code, …
- Relationship to the world:

Form v. meaning v. use v. world

- Form: ഒരു ഭാഷ ഒരിക്കലും മതിയാവില്ല
- Conventional meaning: “One language is never enough”
- Speaker intent: at a tourist market, after a long trip, while maintaining legacy code, …
- Relationship to the world: priceless!

Thought Experiment 1: Java

Thought Experiment 1: Java

- Model: Any model type at all

Thought Experiment 1: Java

- Model: Any model type at all
    - For current purposes: BERT (Devlin et al 2019), GPT-2 (Radford et al 2019), or similar

Thought Experiment 1: Java

- Model: Any model type at all
    - For current purposes: BERT (Devlin et al 2019), GPT-2 (Radford et al 2019), or similar
- Training data: All well-formed Java code on Github, but only the text of the code

Thought Experiment 1: Java

- Model: Any model type at all
    - For current purposes: BERT (Devlin et al 2019), GPT-2 (Radford et al 2019), or similar
- Training data: All well-formed Java code on Github, but only the text of the code
- Test input: A single Java program, possibly even from the training data

Thought Experiment 1: Java

- Model: Any model type at all
    - For current purposes: BERT (Devlin et al 2019), GPT-2 (Radford et al 2019), or similar
- Training data: All well-formed Java code on Github, but only the text of the code
- Test input: A single Java program, possibly even from the training data
- Expected output: Result of executing that program

Thought Experiment 2: English

Thought Experiment 2: English

- Model: Any model type at all

Thought Experiment 2: English

- Model: Any model type at all
    - For current purposes: BERT, GPT-2, or similar

Thought Experiment 2: English

- Model: Any model type at all
    - For current purposes: BERT, GPT-2, or similar
- Training data: As much well-formed English text as you like, but no further info

Thought Experiment 2: English

- Model: Any model type at all
    - For current purposes: BERT, GPT-2, or similar
- Training data: As much well-formed English text as you like, but no further info
    - Not arranged into question/answer pairs and marked as such, etc.

Thought Experiment 2: English

- Model: Any model type at all
    - For current purposes: BERT, GPT-2, or similar
- Training data: As much well-formed English text as you like, but no further info
    - Not arranged into question/answer pairs and marked as such, etc.
- Test input: A photograph plus a sentence like *How many dogs are jumping?* or *Kim said “What a cute puppy!” What is cute?*

Thought Experiment 2: English

r

as you like, but no further

marked as such, etc.

- Test input: A photograph plus a sentence like *How many dogs are jumping?* or *Kim said “What a cute puppy!” What is cute?*

Thought Experiment 2: English

- Test input: A photograph plus a sentence like *How many dogs are jumping?* or *Kim said “What a cute puppy!” What is cute?*

Thought Experiment 2: English

- Test input: A photograph plus a sentence like *How many dogs are jumping?* or *Kim said “What a cute puppy!” What is cute?*

- Expected output: *Three* or the region of the photo with the cute puppy.

That’s not fair!

- Of course not! What’s interesting about these thought experiments is what makes the tests unfair

- They’re unfair because the training data is insufficient for the task

- What’s missing: Meaning!

That’s not fair!

- Of course not! What’s interesting about these thought experiments is what makes the tests unfair

- They’re unfair because the training data is insufficient for the task

- What’s missing: Meaning!

You can’t learn meaning from form alone

So what do they learn?

- If the big transformers aren't learning meaning, what makes them so effective?

- The ability to learn patterns:
  - Lexical similarity
  - Structural regularities
  - Artifacts in the data (Niven & Kao 2019)

- Useful, but not meaning and therefore not a path to general-purpose NLU

Adding Meaning to Training Data

- Stars on starred reviews (e.g. Yelp Inc, 2013)
- SQL queries paired with English queries (e.g. Zelle & Mooney, 1996)
- Paragraphs paired with hypotheses and entailment annotations (NLI datasets, e.g. Bowman et al, 2015)
- Photographs annotated with question/answer (VQA; Antol et al 2015)
- Word problems paired with algebraic equations (e.g. Kushman et al 2014)
- Voice assistant commands paired with expected actions
- …

Adding Meaning to Training Data

- Stars on starred reviews (e.g. Yelp Inc, 2013)
- SQL queries
- Paragraphs p
  e.g. Bowman
- Photographs

How much of this is required before a system can learn what *insufficiently spicy* means?

- Word problems paired with algebraic equations (e.g. Kushman et al 2014)
- Voice assistant commands paired with expected actions
- …

Outline

- Why this argument needs to be made
- Form v. meaning v. use v. world
- **Linguistic knowledge (for developers and machines)**
- Leveraging compositionality

Complementary source of knowledge: Linguistics

- How language works
- Structures at varying levels
- How people learn language
- How people use language
- How language varies and changes over time

Linguistics in NLP

- Design of rule-based systems
- Design of annotation schemas to support machine learning
- Feature engineering in (older) machine learning
- Error analysis

Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax

---

#1 Morphosyntax is the difference between a sentence and a bag of words.

#20 Languages vary in how many morphemes they have per word (on average and maximally).

#46 Constraints ruling out some strings as ungrammatical usually also constrain the range of possible semantic interpretations of other strings.

#49 There is no one universal set of parts of speech, even among the major categories.

#88 Many (all?) languages have semantically empty words which serve as syntactic glue.

Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics

---

- with Alex Lascarides; forthcoming 2019

#4 Meaning derived from form is different from meaning in context of use.

#30 Words can have surprising nonce uses through meaning transfer.

#62 Evidentials encode the source a speaker credits the information to and/or the degree of certainty the speaker feels about it.

#76 Reference resolution depends on discourse structure.

#30 Words can have surprising nonce uses through meaning transfer

- Nunberg (2004) argues that it's the predicates in (58a-e), not the arguments that bear transferred meanings

(58) a. We are parked out back.
b. I am parked out back and have been waiting for 15 minutes.
c. *I am parked out back and may not start.
d. Ringo squeezed himself into a narrow space.
e. Yeats did not like to hear himself read in an English accent.
f. The ham sandwich and salad at table 7 is getting impatient.

- (58f) involves a transferred predicate that is part of a noun phrase.

#46 It's challenging to represent the relationship between the meaning of an idiom and the meaning of its parts

- Nunberg et al (1994): Many idioms aren't completely fixed phrases, but interact with internal modification (96a), information structure (96b), pronominal reference (96c), ellipsis (96d), and coordination (96e):

(96) a. The team left no legal stone unturned.
b. Those strings, he wouldn't pull for you.
c. We worried that Pat might spill the beans, but it was Chris who finally spilled them.
d. My goose is cooked, but yours isn't.
e. Reinventing and Tilting At the Federal Windmill

- Riehemann (2001): Distribute meaning of idiom across the words, but idiomatic words are only licensed by semantic constructions which also require the rest of the idiom to be present.

#79: Some linguistic expressions pass embedded presuppositions up, some don't, and with others it depends

- Holes, plugs, and filters (Karttunen 1973)

- Holes:

(239)
a. Kim stopped smoking.
b. Kim didn't stop smoking.
c. Kim hesitated to stop smoking.
d. It surprised Sandy that Kim hesitated to stop smoking.
e. Pat knew that it surprised Sandy that Kim hesitated to stop smoking.

- Plugs:

(240)
a. Kim promised the kids to introduce them to the present king of France.
b. Kim accused the kids of hiding their candy.
c. Kim asked Sandy to read the book again.

#79: Some linguistic expressions pass embedded presuppositions up, some don’t, and with others it depends

- Holes, plugs, and filters (Karttunen 1973)

- Filters:

(241) a. Sandy believes that if the medicine cabinet door is open, then Kim’s cousin took an apsirin.

b. Sandy believes that if Kim has a cousin, then Kim’s cousin took an aspirin.

(242) a. Sandy believes that Kim’s cousin had a headache and Kim’s cousin took an aspirin.

b. Sandy believes that Kim has a cousin and Kim’s cousin took an aspirin.

(243) a. Sandy believes that either the medicine cabinet door is closed, or Kim’s cousin took an aspirin.

b. Sandy believes that either Kim doesn’t have a cousin, or Kim’s cousin took an aspirin.

Why know these things?

- Better understanding of what is being fed into large machine learning models
- Better error analysis of what goes wrong
- Better understanding of the challenges between modern technology and full-scale, task-independent NLU

Outline

- Why this argument needs to be made
- Form v. meaning v. use v. world
- Linguistic knowledge (for developers and machines)
- **Leveraging compositionality**

A meaning representation system is compositional if (working definition; Bender et al 2015):

- it is grounded in a finite (possibly large) number of atomic symbol-meaning pairings
- it is possible to create larger symbol-meaning pairings by combining the atomic pairings through a finite set of rules;
- the meaning of any non-atomic symbol-meaning pairing is a function of its parts and the way they are combined;
- this function is possibly complex, containing special cases for special types of syntactic combination, but only draws on the immediate constituents and any semantic contribution of the rule combining them; and
- further processing will not need to destructively change a meaning representation created in this way to create another of the same type.

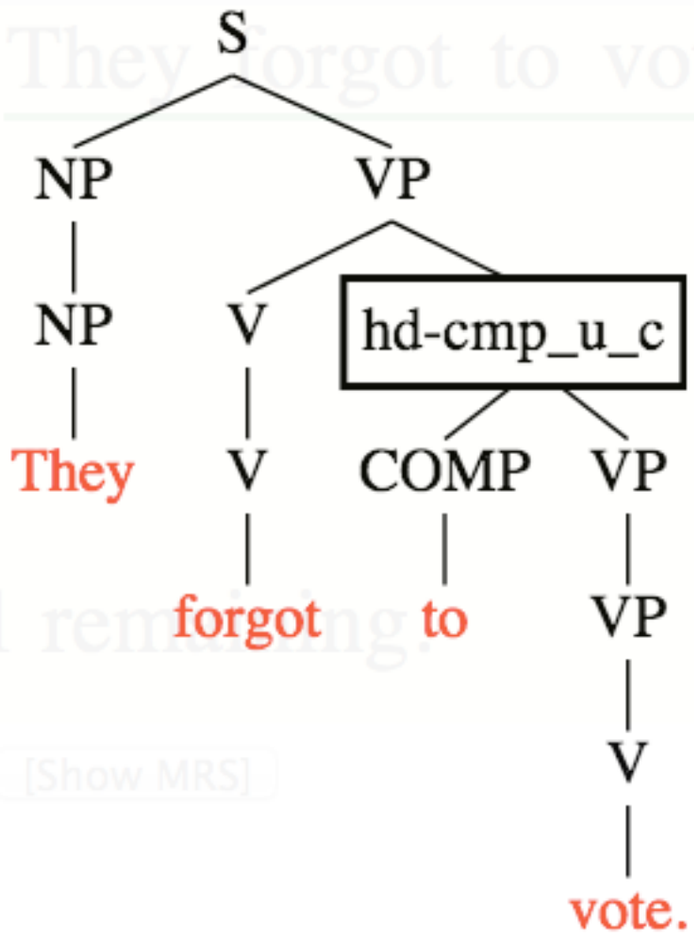
Semantic annotation survey: Compositional layer

- Predicate-argument structure
- Partial constraints on:
  - Scope of negation and other operators
  - Restriction of quantifiers
  - Modality
  - Tense/aspect/mood
  - Information structure
  - Discourse status of referents of NPs
  - Politeness
- Possibly compositional, but not according to sentence grammar:
  - Coherence relations/rhetorical structure

ERG: The English Resource Grammar (Flickinger 2000, 2011)

- Under continuous development since 1993
- Framework: Head-driven Phrase Structure Grammar (Pollard & Sag 1994)
- 1214 release: 225 syntactic rules, 70 lexical rules, 975 leaf lexical types
- Open-source and compatible with open-source DELPH-IN processing engines (www.delph-in.net)
- Broad-coverage: 85-95% on varied domains: newspaper text, Wikipedia, bio-medial research literature (Flickinger et al 2010, 2012; Adolphs et al 2008)
  - Robust processing strategies enable 100% coverage
- Output: derivation trees paired with meaning representations in the Minimal Recursion Semantics framework---English Resource Semantics (ERS)
  - Emerging documentation at moin.delph-in.net/ErgSemantics

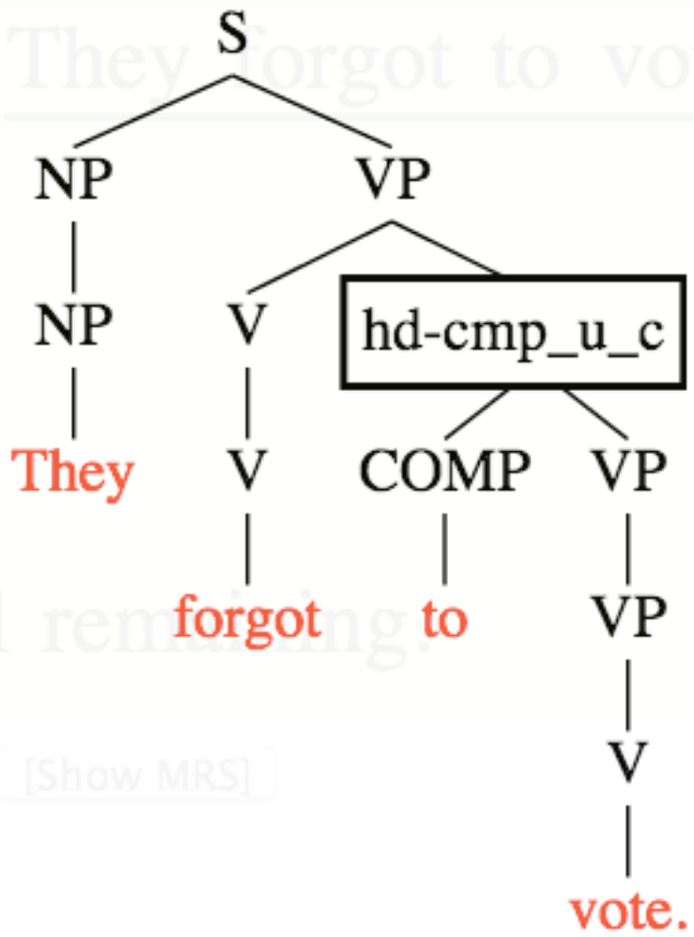
ERG: Examples

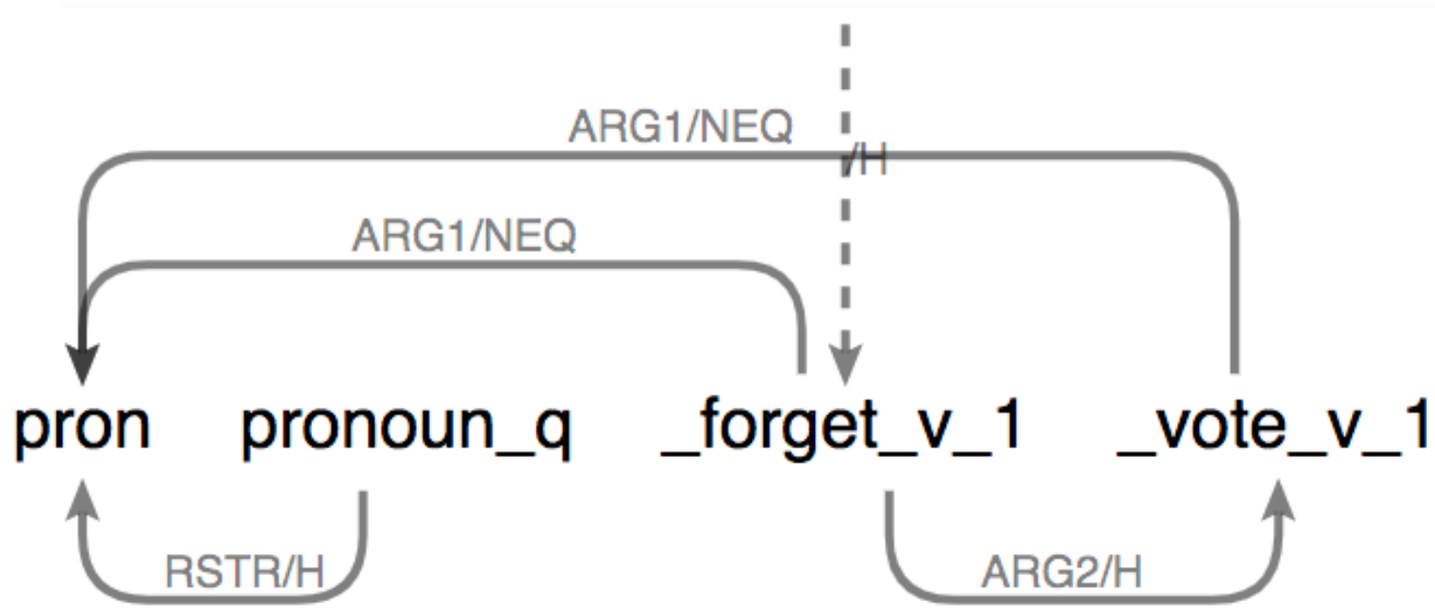


TOP: **h0**
INDEX: **e2**
RELS:
h4:pron_rel(ARG0: **x3**)
h5:pronoun_q_rel(ARG0: **x3**,RSTR: **h6**,BODY: **h7**)
h1:"_forget_v_1_rel"(ARG0: **e2**,ARG1: **x3**,ARG2: **h8**)
h9:"_vote_v_1_rel"(ARG0: **e10**,ARG1: **x3**)

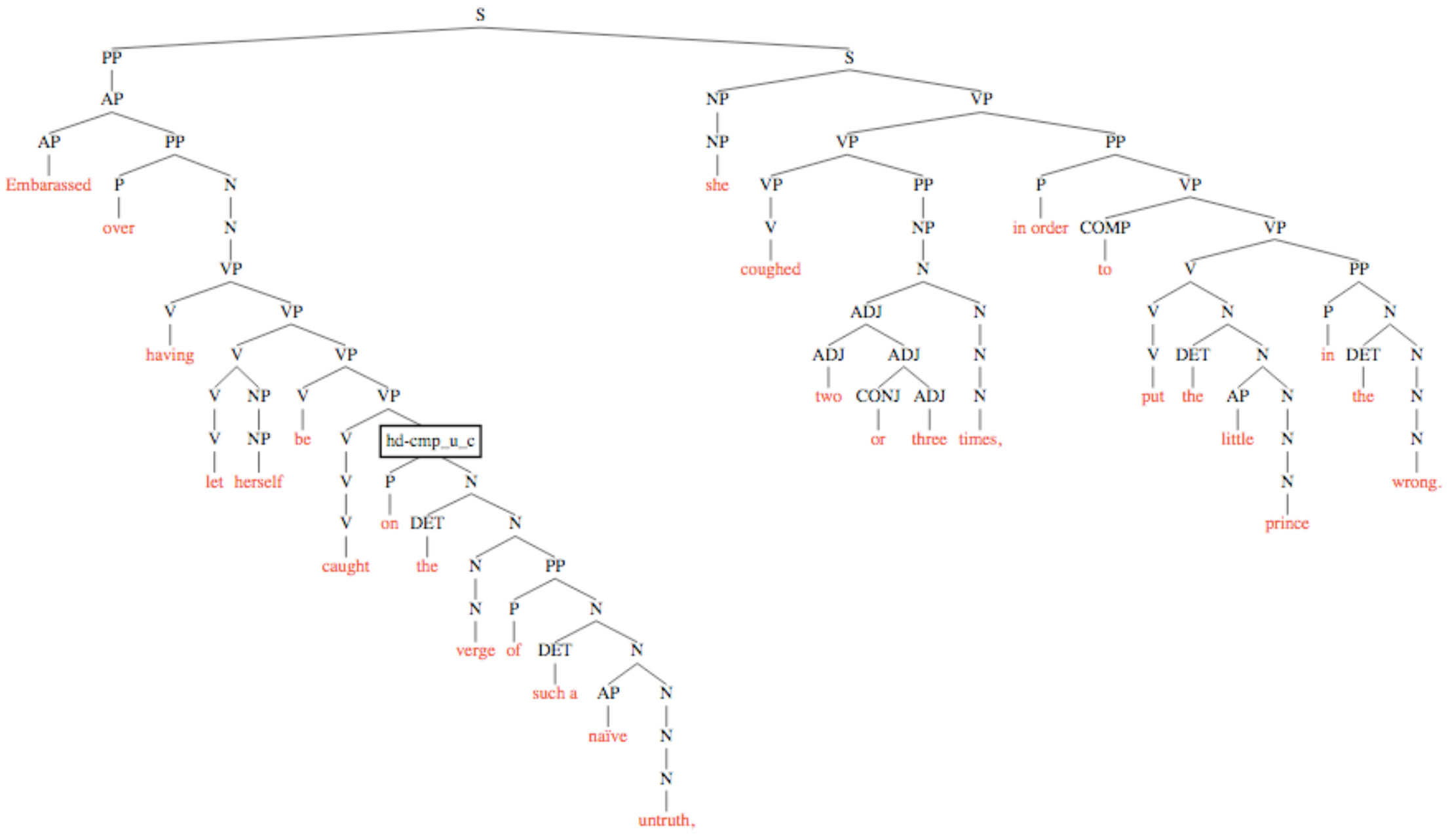
HCONS: **h0** $=_q$ **h1**, **h6** $=_q$ **h4**, **h8** $=_q$ **h9**

ERG: Examples





ERG: Examples



ERG: Examples

INDEX: **e2**
RELS:
h1:subord_rel(ARG0: **e4**,ARG1: **h5**,ARG2: **h6**)
h7:"_embarassed/JJ_u_unknown_rel"(ARG0: **e8**,ARG1: **i9**)
h7:_over_p_rel(ARG0: **e10**,ARG1: **e8**,ARG2: **x11**)
h12:udef_q_rel(ARG0: **x11**,RSTR: **h13**,BODY: **h14**)
h15:nominalization_rel(ARG0: **x11**,ARG1: **h16**)
h16:"_let_v_1_rel"(ARG0: **e17**,ARG1: **i18**,ARG2: **h19**)
h20:pron_rel(ARG0: **x21**)
h22:pronoun_q_rel(ARG0: **x21**,RSTR: **h23**,BODY: **h24**)
h25:"_catch_v_1_rel"(ARG0: **e26**,ARG1: **i27**,ARG2: **x21**,ARG3: **h28**)
h25:parg_d_rel(ARG0: **e29**,ARG1: **e26**,ARG2: **x21**)
h30:_on_p_rel(ARG0: **e31**,ARG1: **x21**,ARG2: **x32**)
h33:_the_q_rel(ARG0: **x32**,RSTR: **h34**,BODY: **h35**)
h36:"_verge_n_1_rel"(ARG0: **x32**)
h36:_of_p_rel(ARG0: **e37**,ARG1: **x32**,ARG2: **x38**)
h39:_such+a_q_rel(ARG0: **x38**,RSTR: **h40**,BODY: **h41**)
h42:"_naïve/JJ_u_unknown_rel"(ARG0: **e43**,ARG1: **x38**)
h42:"_untruth_n_1_rel"(ARG0: **x38**)
h44:pron_rel(ARG0: **x3**)
h45:pronoun_q_rel(ARG0: **x3**,RSTR: **h46**,BODY: **h47**)
h48:"_cough_v_1_rel"(ARG0: **e2**,ARG1: **x3**)
h48:loc_nonsp_rel(ARG0: **e49**,ARG1: **e2**,ARG2: **x50**)
h51:udef_q_rel(ARG0: **x50**,RSTR: **h52**,BODY: **h53**)
h54:card_rel(CARG: "2",ARG0: **e56**,ARG1: **x50**)
h57:_or_c_rel(ARG0: **e58**,L-INDEX: **e56**,R-INDEX: **e59**,L-HNDL: **h54**,R-HNDL: **h60**)
h60:card_rel(CARG: "3",ARG0: **e59**,ARG1: **x50**)
h57:"_times_n_1_rel"(ARG0: **x50**)
h62:"_in+order+to_x_rel"(ARG0: **e63**,ARG1: **h64**,ARG2: **h65**)
h66:"_put_v_1_rel"(ARG0: **e67**,ARG1: **x3**,ARG2: **x68**,ARG3: **h69**)
h70:_the_q_rel(ARG0: **x68**,RSTR: **h71**,BODY: **h72**)
h73:"_little_a_1_rel"(ARG0: **e74**,ARG1: **x68**)
h73:"_prince_n_of_rel"(ARG0: **x68**,ARG1: **i75**)
h76:_in_p_rel(ARG0: **e77**,ARG1: **x68**,ARG2: **x78**)

Redwoods: ERG-based treebanking (sembanking) (Oepen et al 2004)

- Minimal discriminants (Carter 1997): Properties of derivation trees partitioning parse forest per item
- Allows annotators to swiftly navigate even very large parse forests to select intended analysis or reject all analyses
    - 37,200 words of the Brown corpus annotated in 1400 minutes (1.7 sentences/min)
- All annotation decisions are recorded and can be rerun against updated parse forests produced by updated grammar versions
- Current Redwoods release (9th growth) includes 85,000 sentences of annotated text across genres including Wikipedia, tourism brochures, ...

Redwoods: ERG-based treebanking (sembanking) (Oepen et al 2004)

- Analyses can be viewed as full HPSG analyses, ERS only, or even simpler syntactic or semantic dependency representations
- Data source behind
  - ‘DM’ representations at the SDP 2014 and 2015 shared tasks (Oepen et al 2014, 2015) http://sdp.delph-in.net/
  - ‘DM’ and ‘EDS’ representations in the CONLL 2019 shared task http://mrp.nlpl.eu/
- Unlimited ‘silver’ data can be generated at will using the grammar-based parser & treebank trained parse selection model
  - Beneficial in e.g. neural sentence realization (Hajdik et al 2019)

Why a grammar-based compositional approach?

- Importance of task-independent, sentence-meaning annotations
- Can created be done:
  - Non-compositionally, as in Abstract Meaning Representation (AMR; Langkilde & Knight 1998, Banarescu et al 2013)
  - Compositionally, by hand, as in PropBank (Kingsbury & Palmer 2002) and FrameNet (Baker et al 1998)
  - Compositionally, with a machine-readable grammar, as in Redwoods (Oepen et al 2004), TREPIL (Rosén et al 2005), or the Groningen Meaning Bank (Basile et al 2012)

Benefits of compositionality: Comprehensiveness

- Grammar-based compositional approach ⇒ Every word and syntactic structure must be accounted for, or specifically deemed semantically void

- Narrower paraphrase sets, compare AMR (1), (2) (Banarescu et al 2014) to ERS (3)

(1) a. No one ate.
b. Every person failed to eat.

(2) a. The boy is responsible for the work.
b. The boy is responsible for doing the work.
c. The boy has the responsibility for the work.

Benefits of compositionality: Comprehensiveness

- Grammar-based compositional approach ⇒ Every word and syntactic structure must be accounted for, or specifically deemed semantically void

- Narrower paraphrase sets, compare AMR (1), (2) (Banarescu et al 2014) to ERS (3)

(3) a. Kim thinks Sandy gave the book to Pat.
b. Kim thinks that Sandy gave the book to Pat.
c. Kim thinks Sandy gave Pat the book.
d. Kim thinks the book was given to Pat by Sandy.
e. The book, Kim thinks Sandy gave to Pat.

Benefits of compositionality: Comprehensiveness

- Task-independent semantic representations can’t abstract away from seemingly less relevant nuances of sentence meaning

- Compositional approach facilitates capturing more detail

$\langle h_1,$
$\quad h_4$:person$\langle 0{:}6\rangle$(ARG0 x_5),
$\quad h_6$:_no_q$\langle 0{:}6\rangle$(ARG0 x_5, RSTR h_7, BODY h_8),
$\quad h_2$:_eat_v_1$\langle 7{:}11\rangle$(ARG0 e_3, ARG1 x_5, ARG2 i_9)
$\{ h_1 =_q h_2, h_7 =_q h_4 \}\rangle$

```
(e / eat-01
 :polarity -
 :ARG0 (p / person
   :mod (e / every)))
```

$\langle h_1,$
$\quad h_4$:_every_q$\langle 0{:}5\rangle$(ARG0 x_6, RSTR h_7, BODY h_5),
$\quad h_8$:_person_n_1$\langle 6{:}12\rangle$(ARG0 x_6),
$\quad h_2$:_fail_v_1$\langle 13{:}19\rangle$(ARG0 e_3, ARG1 h_9),
$\quad h_{10}$:_eat_v_1$\langle 23{:}27\rangle$(ARG0 e_{11}, ARG1 x_6, ARG2 i_{12})
$\{ h_1 =_q h_2, h_7 =_q h_8, h_9 =_q h_{10} \}\rangle$

Benefits of Compositionality: Consistency

- Requiring meaning representations to be grounded in both the lexical items and syntactic structure of the strings being annotated significantly reduces the space of possible annotations

- Grammar based approach allows encoding of design decisions for machine application
    - Ex: arguments of *when*

- Human input still required, but choosing among representations is far simpler than authoring them
    - Development of grammar is still a big investment, but with big returns as the same grammar is applied over more and more text

Benefits of Compositionality: Scalability

- In amount of text annotated: Initial development of grammar pays off as it is applied to as much text as desired

- In genre diversity of the resource: One and the same grammar can be applied to texts from multiple different domains

    - Robustness techniques can compensate for lack of grammar coverage

- In the complexity of the annotations themselves: Grammar updates can be efficiently propagated across the treebank by reparsing corpus and rerunning annotation decisions (Oepen et al 2004)

    - Improve analyses of particular phenomena, or add layers of grammar-based annotation (e.g. partial constraints on information structure)

Inter-Annotator Agreement study

- Data source: Sentences sampled from Antoine de Saint Exupéry's *The Little Prince*

- Three expert annotators

- Annotated 50-sentence trial set, then adjudicated, updating annotation guidelines as indicated

- Annotated 150-sentence sample set, then measured IAA, then produced adjudicated gold standard

- Repeat above steps with 'bridging' analyses in

Agreement Metrics

- NB: Chance-corrected IAA measures as yet unavailable for graph-structured annotations

- Exact match: Full ERS identical between annotators

- Elementary Dependency Match (Dridan & Oepen 2011)

  - Computed over sets of triples from reduction of ERS to Elementary Dependency Structures (EDS)

  - EDMa: Argument identification only

  - EDMna: Argument identification + predicate name identification

IAA Results

Metric	Annotator Comparison			
	A vs. B	A vs. C	B vs. C	Average
Exact Match	0.73	0.65	0.70	0.70
EDM_a	0.93	0.92	0.94	0.93
EDM_{na}	0.94	0.94	0.95	0.94

- Compare Banarescu et al (2013) triple-based IAA for AMR over web text of 0.71

This talk in a nutshell

- A whole pile of end-to-end systems does not general-purpose NLU make
- Systems, no matter how complex, trained only on form, won't learn meaning
    - That's not how babies do it either
- General-purpose NLU requires attention to linguistic structure and use
- Compositionality is key!

References

Adolphs, P., Oepen, S., Callmeier, U., Crysmann, B., Flickinger, D., and Kiefer, B. (2008). Some fine points of hybrid natural language parsing. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. (2015). VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 17th Meeting of the Association for Computational Linguistics*, pages 86–90.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2014). Abstract Meaning Representation (AMR) 1.1 specification. Version of February 11, 2014.

Basile, V., Bos, J., Evang, K., and Venhuizen, N. (2012). A platform for collaborative semantic annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 92–96, Avignon, France. Association for Computational Linguistics.

Bender, E. M. (2013). *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. Morgan & Claypool.

Bender, E. M. and Lascarides, A. (forthcoming 2019). *Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics*. Morgan & Claypool.

Bender, E. M., Flickinger, D., Oepen, S., Packard, W., and Copestake, A. (2015). Layers of interpretation: On grammar and compositionality. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 239–249, London, UK. Association for Computational Linguistics.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Bruner, J. (1985). Child's talk: Learning to use language. *Child Language Teaching and Therapy*, **1**(1), 111–114.

Carter, D. (1997). The TreeBanker. A tool for supervised training of parsed corpora. In *Proceedings of the Workshop on Computational Environments for Grammar Development and Linguistic Engineering*, pages 9–15, Madrid, Spain.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dridan, R. and Oepen, S. (2011). Parser evaluation using elementary dependency matching. In *Proceedings of the 12th International Conference on Parsing Technologies*, pages 225–230, Dublin, Ireland.

Flickinger, D. (2000). On building a more efficient grammar by exploiting types. *Natural Language Engineering*, **6**(1 Special Issue on Efficient Processing with HPSG), 15 – 28.

Flickinger, D. (2011). Accuracy v. robustness in grammar engineering. In E. M. Bender and J. E. Arnold, editors, *Language from a Cognitive Perspective: Grammar, Usage and Processing*, pages 31–50. CSLI Publications, Stanford, CA.

Flickinger, D., Oepen, S., and Ytrestl, G. (2010). WikiWoods. Syntacto-semantic annotation for English Wikipedia. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Valletta, Malta.

Flickinger, D., Zhang, Y., and Kordoni, V. (2012). DeepBank. A dynamically annotated treebank of the

Wall Street Journal. In *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories*, pages 85–96, Lisbon, Portugal. Edições Colibri.

Hajdik, V., Buys, J., Goodman, M. W., and Bender, E. M. (2019). Neural text generation from rich semantic representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2259–2266, Minneapolis, Minnesota. Association for Computational Linguistics.

Karttunen, L. (1973). Presuppositions of compound sentences. *Linguistic Inquiry*, **4**(2), 169–193.

Kingsbury, P. and Palmer, M. (2002). From TreeBank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*.

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science*, **10**(1), 110–120.

Kushman, N., Artzi, Y., Zettlemoyer, L., and Barzilay, R. (2014). Learning to automatically solve algebra word problems. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 271–281, Baltimore, Maryland. Association for Computational Linguistics.

Langkilde, I. and Knight, K. (1998). Generation that exploits corpus-based statistical knowledge. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 704–710, Montreal, Quebec, Canada. Association for Computational Linguistics.

Niven, T. and Kao, H.-Y. (2019). Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Nunberg, G., Sag, I. A., and Wasow, T. (1994). Idioms. *Language*, **70**(3), 491–538.

Oepen, S., Flickinger, D., Toutanova, K., and Manning, C. D. (2004). LinGO Redwoods. A rich and dynamic treebank for HPSG. *Journal of Research on Language and Computation*, **2**(4), 575 – 596.

Oepen, S., Kuhlmann, M., Miyao, Y., Zeman, D., Flickinger, D., Hajič, J., Ivanova, A., and Zhang, Y. (2014). SemEval 2014 task 8: Broad-coverage semantic dependency parsing. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 63–72, Dublin, Ireland. Association for Computational Linguistics.

Oepen, S., Kuhlmann, M., Miyao, Y., Zeman, D., Cinková, S., Flickinger, D., Hajič, J., and Urešová, Z. (2015). SemEval 2015 task 18: Broad-coverage semantic dependency parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 915–926, Denver, Colorado. Association for Computational Linguistics.

Pollard, C. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. The University of Chicago Press and CSLI Publications, Chicago, IL and Stanford, CA.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. Unpublished MS, OpenAI San Francisco.

Riehemann, S. (2001). *A Constructional Approach to Idioms and Word Formation*. Ph.D. thesis, Stanford University.

Rosén, V., Meurer, P., and Smedt, K. D. (2007). Designing and implementing discriminants for LFG grammars. In *Proceedings of LFG07*, pages 397–417. CSLI On-line Publications.

Tomasello, M. and Farrar, M. J. (1986). Joint attention and early language. *Child Development*, **57**(6), 1454–1463.

Yelp Inc (2013). Yelp dataset challenge.

Zelle, J. M. and Mooney, R. J. (1996). Learning to parse database queries using inductive logic programming. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1050–1055.