Natural Language Inference

Функциональные модели в естественном языке

Natural Language Inference

Natural Language Inference (NLI) -- задача определения логических отношений между словами, фразами, предложениями, параграфами, документами…

Простые примеры

Premise (Посылка)	Relation (Отношение)	Hypothesis (Гипотеза)
elephant		linguist
An elephant danced		An elephant moved
Every reptile danced		An elephant ate
Some elephants walk		No elephants move
James Byron Dean refused to move without blue jeans		James Dean didn’t dance without pants
Mitsubishi Motors Corp’s new vehicle sales in the US fell 46 percent in June		Mitsubishi’s sales rose 46 percent
Acme Corporation reported that its CEO resigned		Acme’s CEO resigned

NLI ставит задачей выявление (логических) отношений между словами (например, elephant -- linguist) и пропозициями (например, some students smoke – some students smoke cigars). В самом простом случае, отношения могут сводиться к следствию (entailment), противоречию (contradiction), отсутствию какого-либо отношения (neutral).

Простые примеры

Premise (Посылка)	**Relation (Отношение)**	**Hypothesis (Гипотеза)**
elephant	contradicts	linguist
An elephant danced	entails	An elephant moved
Every reptile danced	neutral	An elephant ate
Some elephants walk	contradicts	No elephants move
James Byron Dean refused to move without blue jeans	entails	James Dean didn’t dance without pants
Mitsubishi Motors Corp’s new vehicle sales in the US fell 46 percent in June	contradicts	Mitsubishi’s sales rose 46 percent
Acme Corporation reported that its CEO resigned	entails	Acme’s CEO resigned

NLI: формулировка задачи

Можно ли вывести из посылки (premise) гипотезу (hypothesis)?

- ★ Опора на здравый смысл, а не на формальную логику
- ★ Фокус на локальных выводах, а не на цепочках логических выводов
- ★ Упор на разнообразие лингвистических способов выражения логических отношений

Зачем нужен NLI?

Dagan et al. (2006)

It seems that major inferences, as needed by multiple applications, can indeed be cast in terms of textual entailment. For example, **a QA system** has to identify texts that entail a hypothesized answer. [...] Similarly, for certain **Information Retrieval** queries the combination of semantic concepts and relations denoted by the query should be entailed from relevant retrieved documents. [...] In **multi-document summarization** a redundant sentence, to be omitted from the summary, should be entailed from other sentences in the summary. And in **MT evaluation** a correct translation should be semantically equivalent to the gold standard translation, and thus both translations should entail each other. Consequently, we hypothesize that textual entailment recognition is a suitable generic task for evaluating and comparing applied semantic inference models. Eventually, such efforts can promote the development of entailment recognition "engines" which may provide useful generic modules across applications.

Связь с другими задачами

Задача	**В терминах NLI**
Paraphrase	text ≡ paraphrase
Summarization	text $\sqsubset$ summary
Information retrieval	query $\sqsupset$ document
Question answering	question $\sqsupset$ answer Who left?⇒Someone left Someone left $\sqsupset$ Dasha left

Лейблы

couch **sofa**	**crow** **bird**	**bird** **crow**	**hippo** **hungry**	**elephant** **linguist**
yes entailment		**no** non-entailment		
yes entailment		**unknown** non-entailment		**no** contradiction
P ≡ Q equivalence	**P ⊏ Q** forward	**P ⊐ Q** reverse	**P # Q** non-entailment	

Лейблы

Обычно пространство отношений разбивается на две (следствие, не-следствие), три (следствие, не-следствие, противоречие), или четыре (тождественность, включение, обратное включение, не-следствие) категории. Какой из дизайнов кажется вам наиболее перспективным? Что вы думаете по поводу расширения наименований отношений, например, наподобие инвентаря лексических функций?

The Stanford Natural Language Inference (SNLI) Corpus

1. Bowman et al. 2015
2. Все посылки (premises) подписи из корпуса Flickr30K (Young et al. 2014).
3. Все гипотезы (hypotheses) получены с помощью краудсорсинга.
4. 550,152 train examples; 10K dev; 10K test
5. Средняя длина в токенах:
    a. Посылка (premise): 14.1
    b. Гипотеза (hypothesis): 8.3
6. Типы клауз:
    a. Посылка в главном предложении: 74%
    b. Гипотеза в главном предложении: 88.9%
7. Размер словаря: 37,026
8. 56,951 примеров были аннотированы дополнительными аннотаторами:
    a. 58.3% примеров получили ожидаемый лейбл отношения (gold label)
    b. 91.2% ожидаемых лейблов (gold labels) совпадают с ожидаемыми лейблами автора
    c. 0.70 общий индекс согласия между аннотаторами (Fleiss kappa)
9. https://nlp.stanford.edu/projects/snli/

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** an **false** description of the photo.

Photo caption **A little boy in an apron helps his mother cook.**

Definitely correct Example: For the caption *"Two dogs are running through a field."* you could write *"There are animals outdoors."*

Write a sentence that follows from the given caption.

Maybe correct Example: For the caption *"Two dogs are running through a field."* you could write *"Some puppies are running to catch a stick."*

Write a sentence which may be true given the caption, and may not be.

Definitely incorrect Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."*

Write a sentence which contradicts the caption.

Problems (optional) *If something is wrong with the caption that makes it difficult to understand, do your best above and let us know here.*

Примеры

Premise (Посылка)	Relation (Отношение)	Hypothesis (Гипотеза)
A man inspects the uniform of a figure in some East Asian country		The man is sleeping
An older and younger man smiling		Two men are smiling and laughing at the cats playing on the floor
A black race car starts up in front of a crowd of people		A man is driving down a lonely road
A soccer game with multiple males playing		Some men are playing a sport
A smiling costumed woman is holding an umbrella		A happy woman in a fairycostume holds an umbrella

Примеры

Premise (Посылка)	Relation (Отношение)	Hypothesis (Гипотеза)
A man inspects the uniform of a figure in some East Asian country	**contradiction** c c c c c	The man is sleeping
An older and younger man smiling	**neutral** n n e n n	Two men are smiling and laughing at the cats playing on the floor
A black race car starts up in front of a crowd of people	**contradiction** c c c c c	A man is driving down a lonely road
A soccer game with multiple males playing	**entailment** e e e e e	Some men are playing a sport
A smiling costumed woman is holding an umbrella	**neutral** n n e c n	A happy woman in a fairycostume holds an umbrella

Кореференция событий

Premise (Посылка)	Relation (Отношение)	Hypothesis (Гипотеза)
A boat sank in the Pacific Ocean.	**contradiction**	A boat sank in the Atlantic Ocean.
Ruth Bader Ginsburg was appointed to the Supreme Court.	**contradiction**	I had a sandwich for lunch today.

Если посылка и гипотеза *скорее всего* описывают разные фото, то лейбл -- противоречие

The Multi-Genre Natural Language Inference (MultiNLI) corpus

1. Williams et al. 2018
2. Train premises drawn from five genres:
    a. Fiction: works from 1912–2010 spanning many genres
    b. Government: reports, letters, speeches, etc., from government websites
    c. The Slate website
    d. Telephone: the Switchboard corpus
    e. Travel: Berlitz travel guides
3. Additional genres just for dev and test (the mismatched condition):
    a. The 9/11 report
    b. Face-to-face: The Charlotte Narrative and Conversation Collection
    c. Fundraising letters
    d. Non-fiction from Oxford University Press
    e. Verbatim: articles about linguistics
4. 392,702 train examples; 20K dev; 20K test
5. 19,647 examples validated by four additional annotators
    a. 58.2% examples with unanimous gold label
    b. 92.6% of gold labels match the author's label
6. Project page: https://cims.nyu.edu/~sbowman/multinli/

The Cross-lingual Natural Language Inference (XNLI) corpus

- ★ A crowd-sourced collection of 5,000 test and 2,500 dev pairs for the MultiNLI corpus
- ★ The pairs are annotated with textual entailment and translated into 14 languages: *French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili and Urdu*
- ★ This results in 112.5k annotated pairs
- ★ https://cims.nyu.edu/~sbowman/xnli/

Veridicality Assessment

de Marneffe et al. 2012

Что такое веридикальность?

A lexical item **L** is **veridical** if the meaning of **L** applied to argument **p** entails **the truth of p:**

(1) *Mary knows that Ivan smokes cigars => Ivan smokes cigars*

know is a **veridical** predicate

(2) *Mary believes that Ivan smokes cigars !=> Ivan smokes cigars*

believe is a **non-veridical** predicate

- ★ The goal is to begin to identify the linguistic and contextual factors that shape readers' veridicality judgments

- ★ The idea: lexical approaches to veridicality are only part of the bigger mechanism, and veridicality should be assessed using information from the entire sentence as well as from the context

- ★ For example, veridicality of (3) depends on whether we trust CNN:

  (3) *CNN reports that Prince Harry has returned to the US.*

The FactBank

- ★ The **FactBank** corpus is a leading resource for research on veridicality
- ★ It provides veridicality annotations for events relative to each participant involved in the discourse
- ★ Each tag consists of a **veridicality** value (certain [CT], probable [PR], possible [PS], underspecified [U])
- ★ and a **polarity** value (positive [+], negative [−], unknown [u])
- ★ CT+ corresponds to the standard notion of veridicality,
- ★ CT− to anti-veridicality, and Uu to non-veridicality.

The FactBank

Table 1
FactBank annotation scheme. CT = certain; PR = probable; PS = possible; U = underspecified; + = positive; − = negative; u = unknown.

Value	Definition	Count
CT+	According to the source, it is certainly the case that X	7,749 (57.6%)
PR+	According to the source, it is probably the case that X	363 (2.7%)
PS+	According to the source, it is possibly the case that X	226 (1.7%)
CT−	According to the source, it is certainly not the case that X	433 (3.2%)
PR−	According to the source it is probably not the case that X	56 (0.4%)
PS−	According to the source it is possibly not the case that X	14 (0.1%)
CTu	The source knows whether it is the case that X or that not X	12 (0.1%)
Uu	The source does not know what the factual status of the event is, or does not commit to it	4,607 (34.2%)
		13,460

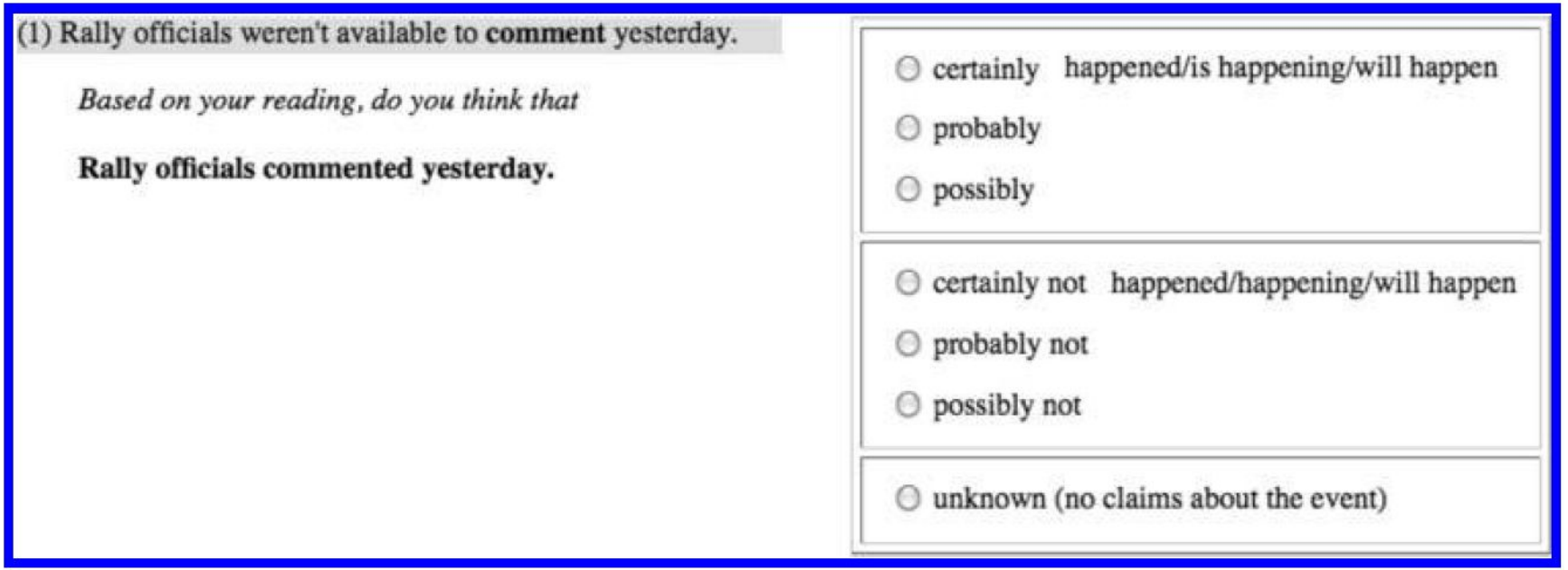
Central claims

- ★ Pragmatically informed veridicality judgments are systematic enough to be included in computational work on textual understanding
- ★ The authors seek to justify FactBank's seven-point categorization over simpler alternatives (e.g., certain vs. uncertain, as in the CoNLL task)
- ★ The inherent uncertainty of pragmatic inference suggests to them that veridicality judgments are not always categorical, and thus are better modeled as probability distributions over veridicality categories

Annotations from the Reader's Perspective

- ★ FactBank seeks to capture aspects of sentence meaning, whereas authors aim to capture aspects of utterance meaning
- ★ They extend the FactBank annotations by bringing world knowledge into the picture
- ★ They use a subset of the FactBank sentences annotated from the author's perspective and recruit MTurkers to annotate the sentences from the reader's perspective

Design of the Mechanical Turk experiment



Annotations from the Reader's Perspective: results

- ★ Authors collected 10 annotations for each of the 642 events, a total of 177 Turkers participated in the annotations.
- ★ At least 6 out of 10 Turkers agreed on the same tag for 500 of the 642 sentences (78%)
- ★ For 53% of the examples, at least 8 Turkers agreed with each other
- ★ Total agreement is obtained for 26% of the data (165 sentences)

An Alternative Scale

- ★One of the goals is to assess whether FactBank's seven-category scheme is the right one for the task
- ★To this end, authors also evaluated whether a five-tag version would increase agreement and perhaps provide a better match with readers' intuitions
- ★Logically, PR− is equivalent to PS+, and PS− to PR+, so it seemed natural to try to collapse them into a two-way division between "probable" and "possible"

An Alternative Scale: results

- ★ The five-point scheme led to lower agreement between Turkers
- ★ Globally, the PR− items were generally mapped to “no”, and PS− to either “no” or “unknown”
- ★ Some Turkers chose the expected mappings (PS− to “probable” and PR− to “possible”), but only very rarely
- ★ Authors conclude from this that Saurı’s 7-point-scale comes closer than its competitors to capturing reader intuitions about veridicality

Lessons from the New Annotations

- ★ Although mostly authors' experiment went in line with FactBank, some differences emerge due to pragmatic enrichment
- ★ For example, *say*, *report*, and *indicate* are tagged Uu in the FactBank
- ★ However, in the PragBank veridicality is more sensitive to context:

  (4) *In the air, U.S. Air Force fliers say they have engaged in "a little cat and mouse" with Iraqi warplanes.*
- ★ This was estimated as certain by 9 of 10 Turkers

Lessons from the New Annotations

- ★ Another difference between the FactBank and the PragBank is the more nuanced categories for PS and PR events
- ★ In FactBank, markers of possibility or probability, such as *could* or *likely*, uniquely determine the corresponding tag
- ★ In contrast, the Turkers allow the bias created by these lexical items to be swayed by other factors

(5) *Iraq could start hostilities with Israel either through a direct attack or by attacking Jordan.*

Annotations: Uu: 6, PS+: 3, PR+: 1

Some unexpected results

- ★ Some lexical markers behave as linguistic theories predict
- ★ For example, *believe* is often a marker of probability whereas *could* and *may* are more likely to indicate possibility
- ★ However, sometime pragmatics corrects it
- ★ The greatest departure from theoretical predictions occurs with the SAY category, which is logically non-veridical but correlates highly with certainty (CT+) in the new corpus
- ★ Conversely, the class KNOW, which includes *know*, *acknowledge*, and *learn*, is traditionally analyzed as veridical (CT+), but in the data is sometimes a marker of possibility

The CommitmentBank

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser (2019). The CommitmentBank: Investigating projection in naturally occurring discourse. Proceedings of Sinn und Bedeutung 23.

The Data: https://github.com/mcdm/CommitmentBank

The CommitmentBank is a corpus of 1,200 naturally occurring discourses whose final sentence contains a clause-embedding predicate under an entailment canceling operator (question, modal, negation, antecedent of conditional)

Jane knows that it is snowing: *know* entails its complement, hence the speaker is committed to the complement clause (CC)

Content which is expressed under the scope of an entailment canceling operator but which is nonetheless understood to be a commitment of the speaker is said to *project*:

Jane doesn't know that it is snowing.

Does Jane know that it is snowing?

Jane may know that it is snowing.

If Jane knows that it is snowing, she will wear her snow boots, hat and gloves.

The goal with the CommitmentBank has been to create a resource for the empirically-based study of projection of CCs, using naturally occurring examples and basing analysis on judgments of projection provided by theoretically untrained speakers.

The CommitmentBank contains 1,200 examples of naturally occurring discourse segments extracted from three corpora of different genres: the Wall Street Journal (WSJ, news articles),the fiction component of the British National Corpus (BNC, fiction) and Switchboard (SWBD,dialogue).

Each discourse consists of a target sentence with a clause-embedding predicate embedded under an entailment canceling operator (negation, modal, antecedent of conditional, or question) with up to 2 prior context sentences/turns.

What fun to hear Artemis laugh. She's such a serious child.
I didn't know she had a sense of humor. [BNC-1607]

A: Oh yes. Animals have a way of talking.
B: Alfie did. I tell you if I could have gotten a hold of that cat that day.
A: I don't know uh that I'd trade my dog in for the world. [SWBD-243]

Predicate	Conditional	Modal	Negation	Question	Predicate	Conditional	Modal	Negation	Question	Predicate	Conditional	Modal	Negation	Question
accept	0	0	1	1	forget	0	4	7	2	recognize	0	0	1	0
admit	1	3	1	1	guarantee	0	2	0	0	remember	1	4	2	0
announce	1	1	0	1	guess	0	6	9	5	see	1	27	10	1
assume	1	7	1	2	hear	2	5	3	1	seem	0	0	2	0
believe	5	19	40	10	hope	0	17	1	2	say	21	40	39	14
bet	0	0	0	1	hypothesize	0	0	0	1	show	0	2	0	0
bother	0	0	1	0	imagine	2	14	12	1	signal	0	1	0	1
convince	0	3	4	0	insist	2	1	0	2	specify	0	1	0	0
decide	3	8	0	0	know	18	16	78	21	suggest	4	3	11	1
demand	0	1	1	0	learn	2	0	2	2	suppose	2	3	2	1
expect	0	1	4	0	mean	4	14	27	7	suspect	5	11	4	0
fear	2	1	0	0	notice	1	7	23	3	swear	0	1	0	0
feel	4	8	16	6	occur	0	0	1	0	take	0	0	0	1
figure	1	0	0	0	pretend	0	2	2	2	tell	6	21	7	4
find	6	9	1	5	prove	0	4	1	0	think	21	39	265	61
foresee	1	0	0	0	realize	0	3	20	6	understand	0	4	4	1

Table 1: Number of discourses by predicate in each embedding environment.

Sally: While the rest of the gang dived into the pub opposite to use the toilets, I called in at H. R. Higgins (Coffee-man) Ltd and bought six gift boxes of coffee (assorted) and two of tea (scented). That was my Christmas shopping sewn up. Who said it was stressful?

Tell us how certain Sally is that Christmas shopping was stressful.

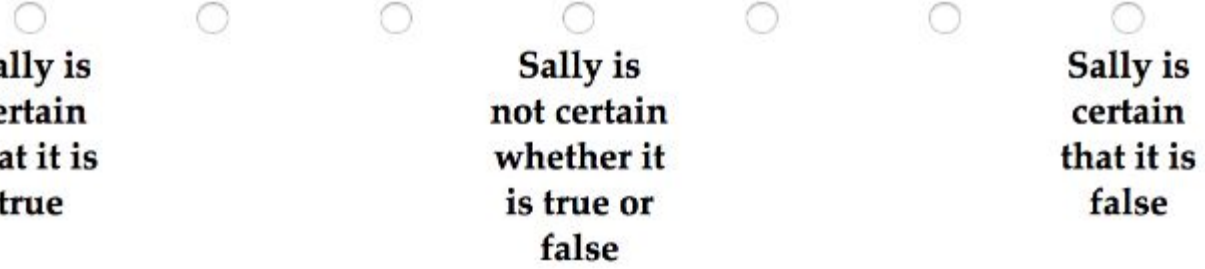


Figure 1: Item display for projection annotation on Mechanical Turk.

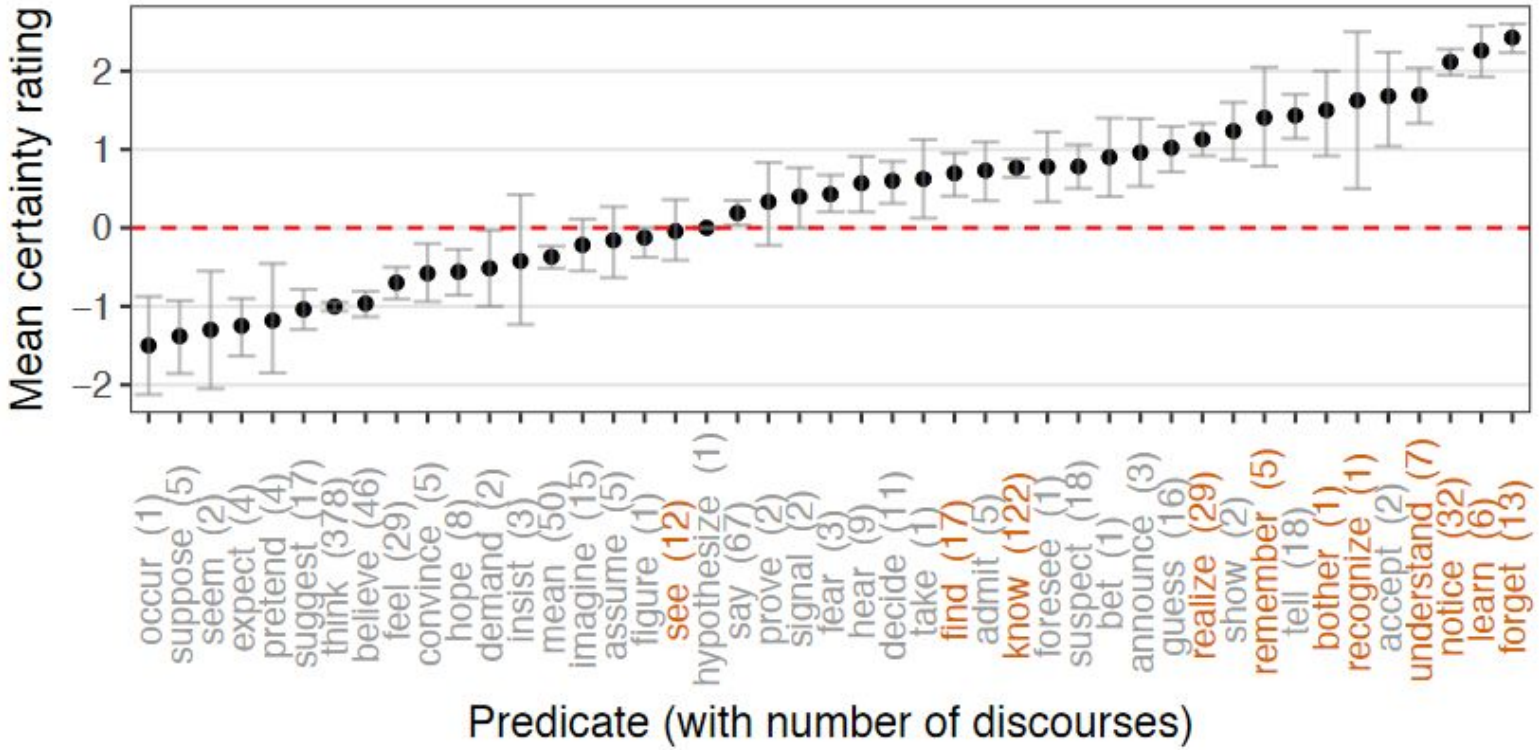
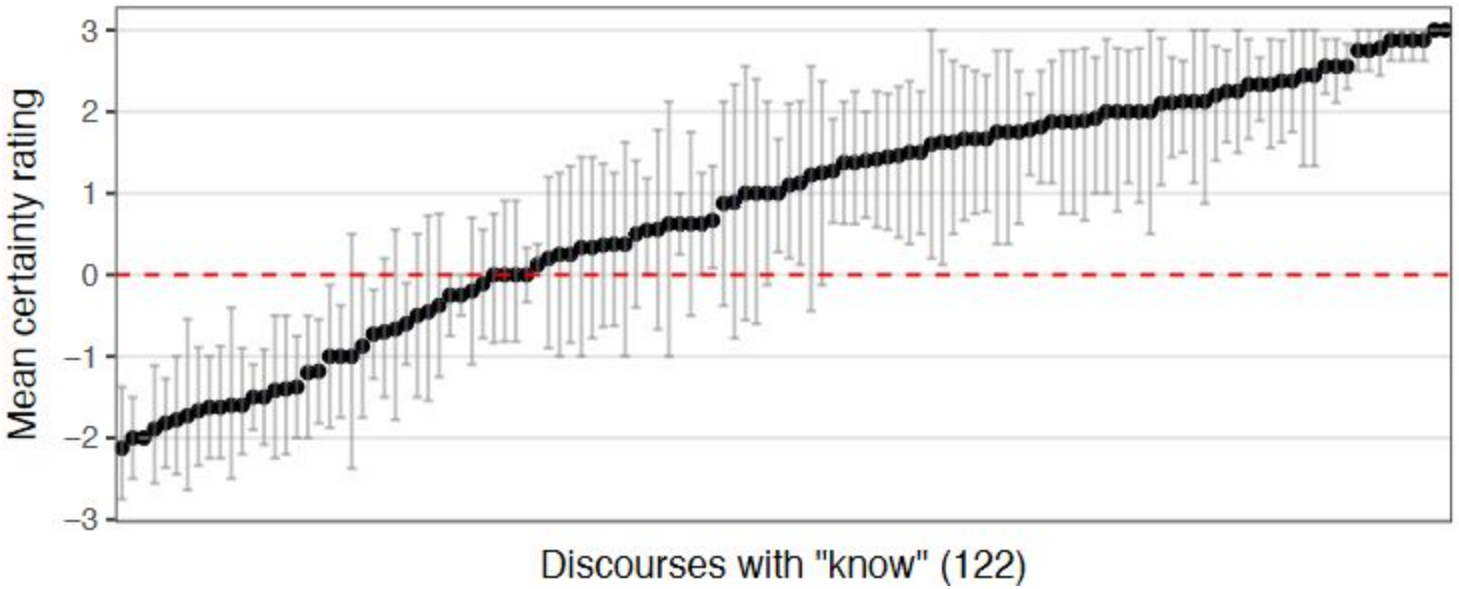
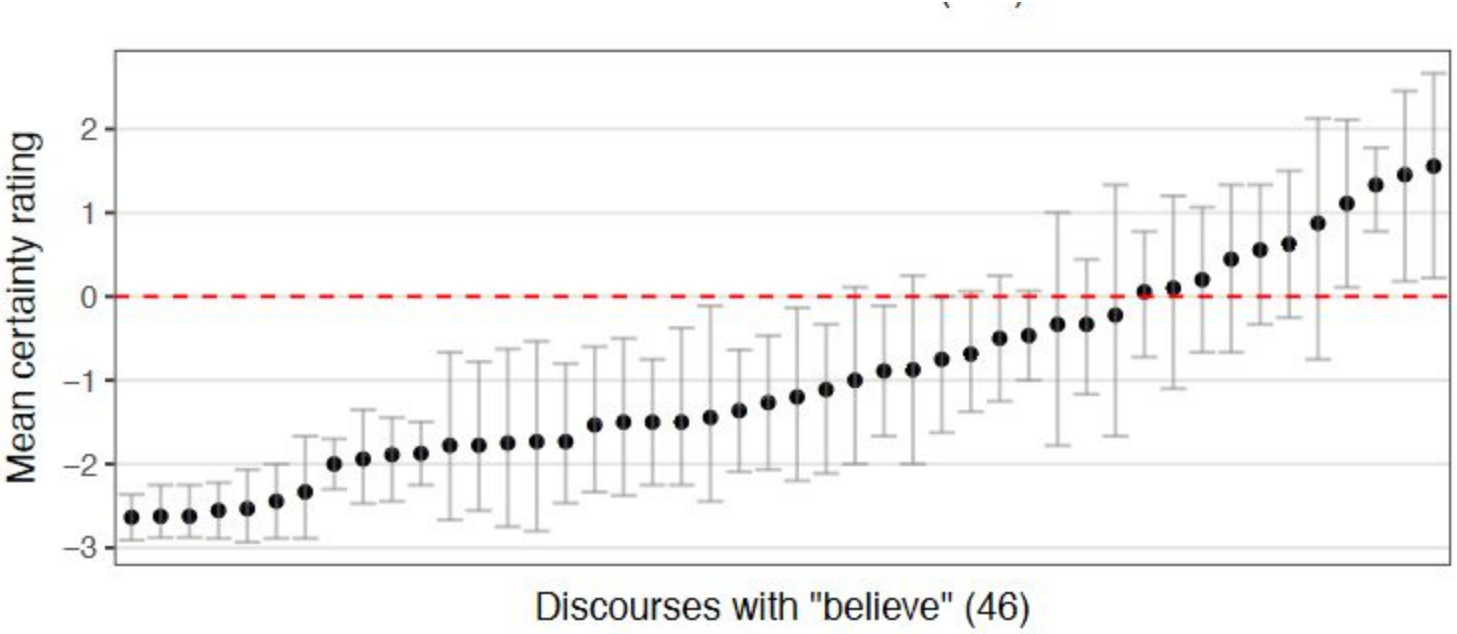
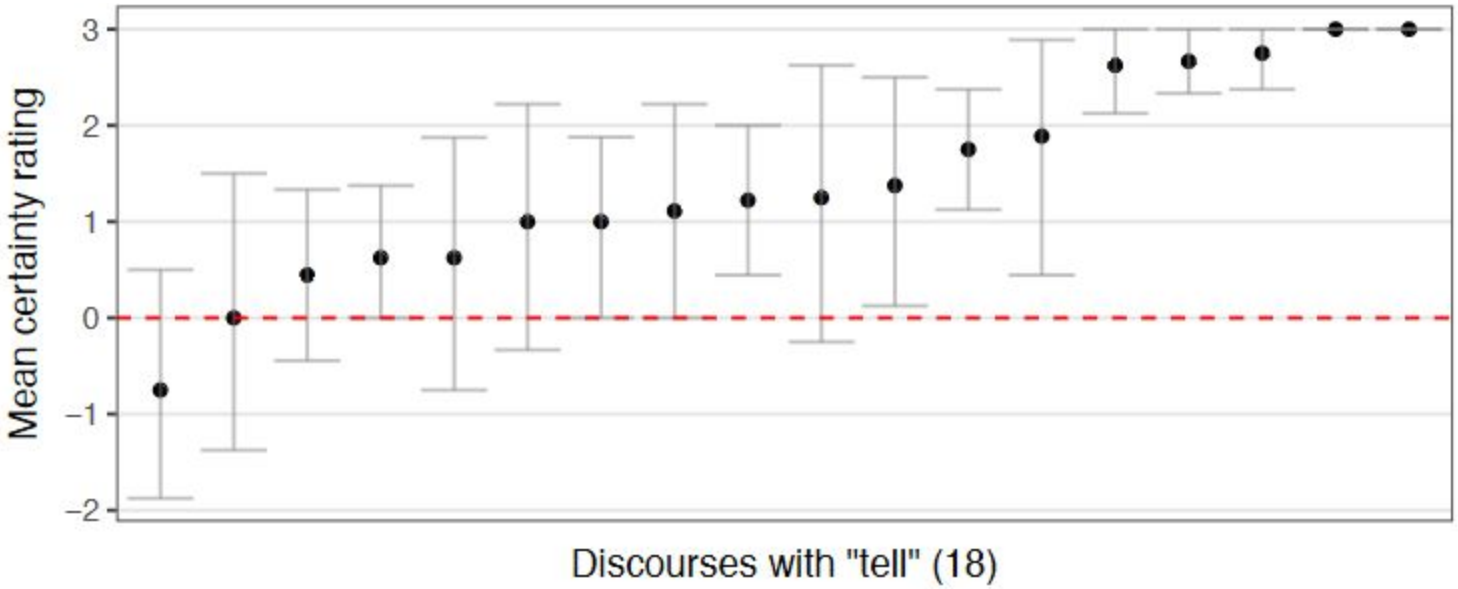


Figure 2: Mean certainty ratings for CCs, by predicate, restricting modal environment to epistemic modals. Number of discourses in parentheses. Error bars indicate bootstrapped 95% confidence intervals. Purported factive predicates are in orange, nonfactive predicates in gray.


Mean certainty rating
3
2
1
0
−1
−2
−3
Discourses with "know" (122)


Mean certainty rating
2
1
0
−1
−2
−3
Discourses with "believe" (46)


Mean certainty rating
3
2
1
0
−1
−2
Discourses with "tell" (18)

Projective or not?

a. At the heart of the universe there is cruelty. We are predators and are preyed upon every living thing. Did you know that wasps lay their eggs in ladybirds piercing the weak spot in their armour? [BNC-2375, mean: +3]

b. "Rather a long shot wasn't it? Twenty years? How do you know the baby was born here?" [BNC-2394, mean: -0.25]

c. The Susweca. It means "dragonfly" in Sioux you know. Did I ever tell you that's where Paul and I met? [BNC-2630, mean: +3]

d. His reaction to the news had been partly predictable and partly complex and more disturbing. There had been the natural initial shock of disbelief at hearing of the unexpected death of any person even casually known. He would have felt no less if he'd been told that Berowne was dead of a coronary or killed in a car smash. [BNC-428, mean: -0.75]

The CommitmentBank can be recast as a dataset for NLI:
https://www.aclweb.org/anthology/D19-1630.pdf