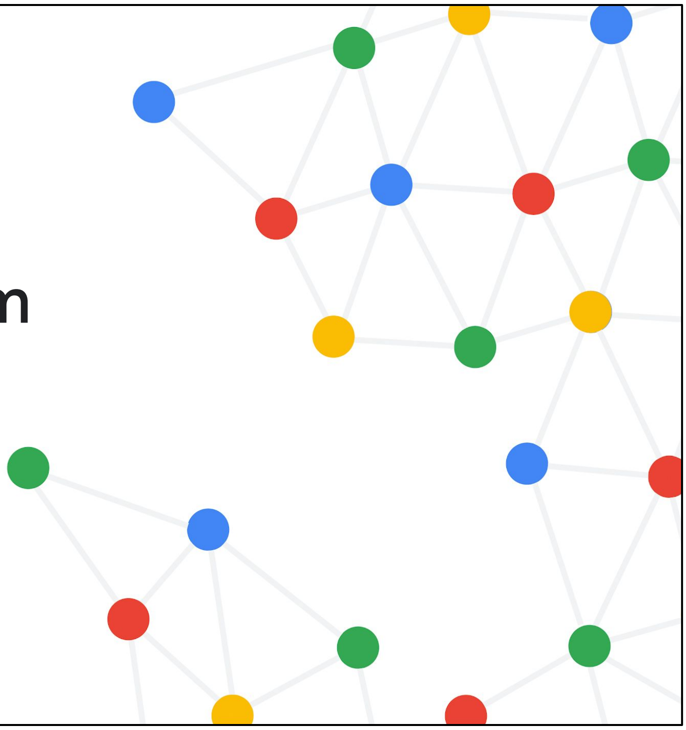
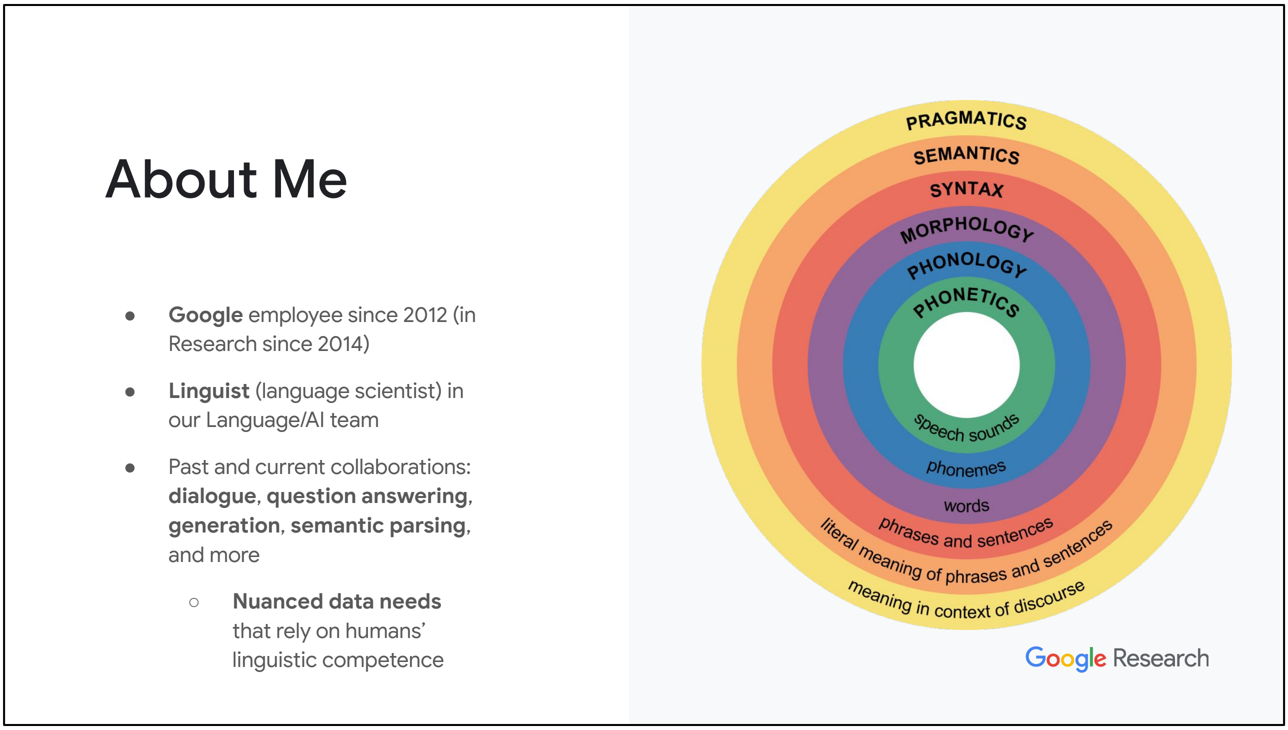
Linguistic Wisdom From the Crowd

Michael Tseng



Google Research



About Me

- **Google** employee since 2012 (in Research since 2014)
- **Linguist** (language scientist) in our Language/AI team
- Past and current collaborations: **dialogue**, **question answering**, **generation**, **semantic parsing**, and more
  - ○ **Nuanced data needs** that rely on humans' linguistic competence

PRAGMATICS
SEMANTICS
SYNTAX
MORPHOLOGY
PHONOLOGY
PHONETICS
speech sounds
phonemes
words
phrases and sentences
literal meaning of phrases and sentences
meaning in context of discourse
Google Research

Image credits:

- Diagram is public domain: https://commons.wikimedia.org/wiki/File:Major_levels_of_linguistic_structure.svg

Agenda

01 Crowdsourcing as Design

02 Linguistic Crowdsourcing

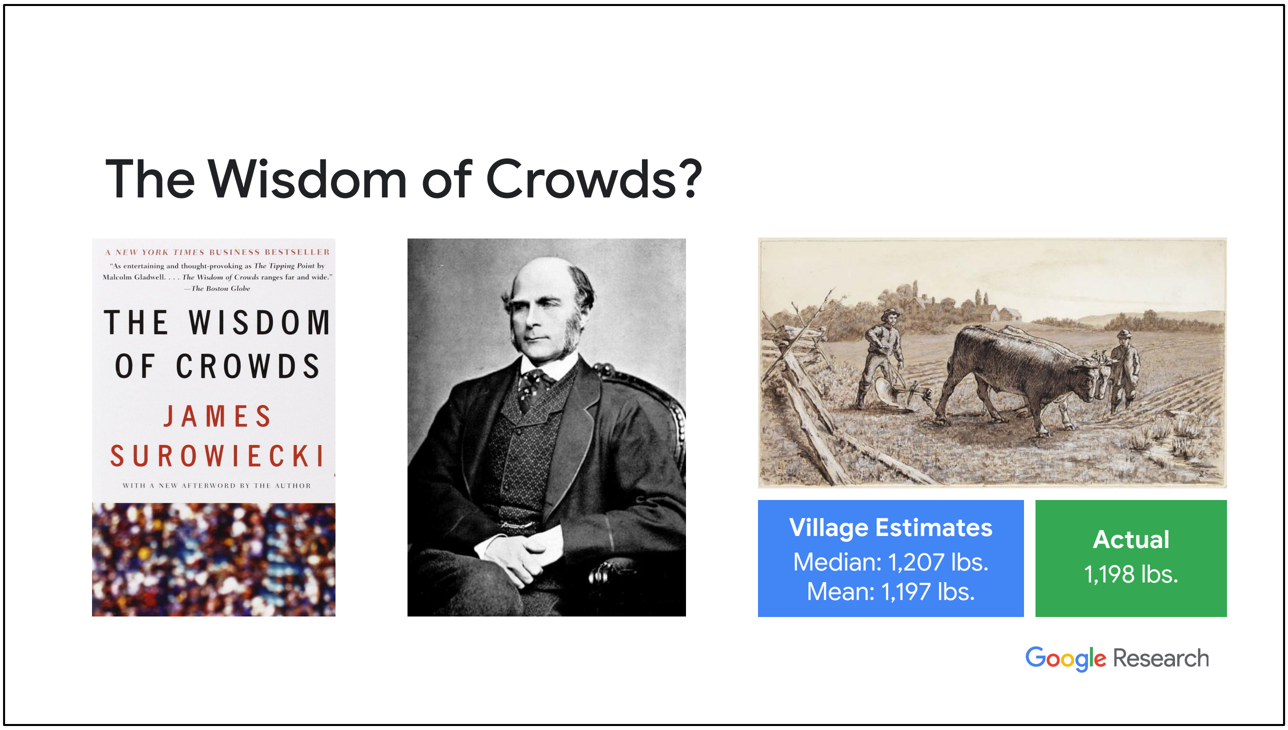
03 Selected Domains of Interest

Google Research

01

Crowdsourcing as Design

Google Research


A NEW YORK TIMES BUSINESS BESTSELLER
THE WISDOM OF CROWDS
JAMES SUROWIECKI
WITH A NEW AFTERWORD BY THE AUTHOR

The Wisdom of Crowds?

Village Estimates
Median: 1,207 lbs.
Mean: 1,197 lbs.

Actual
1,198 lbs.

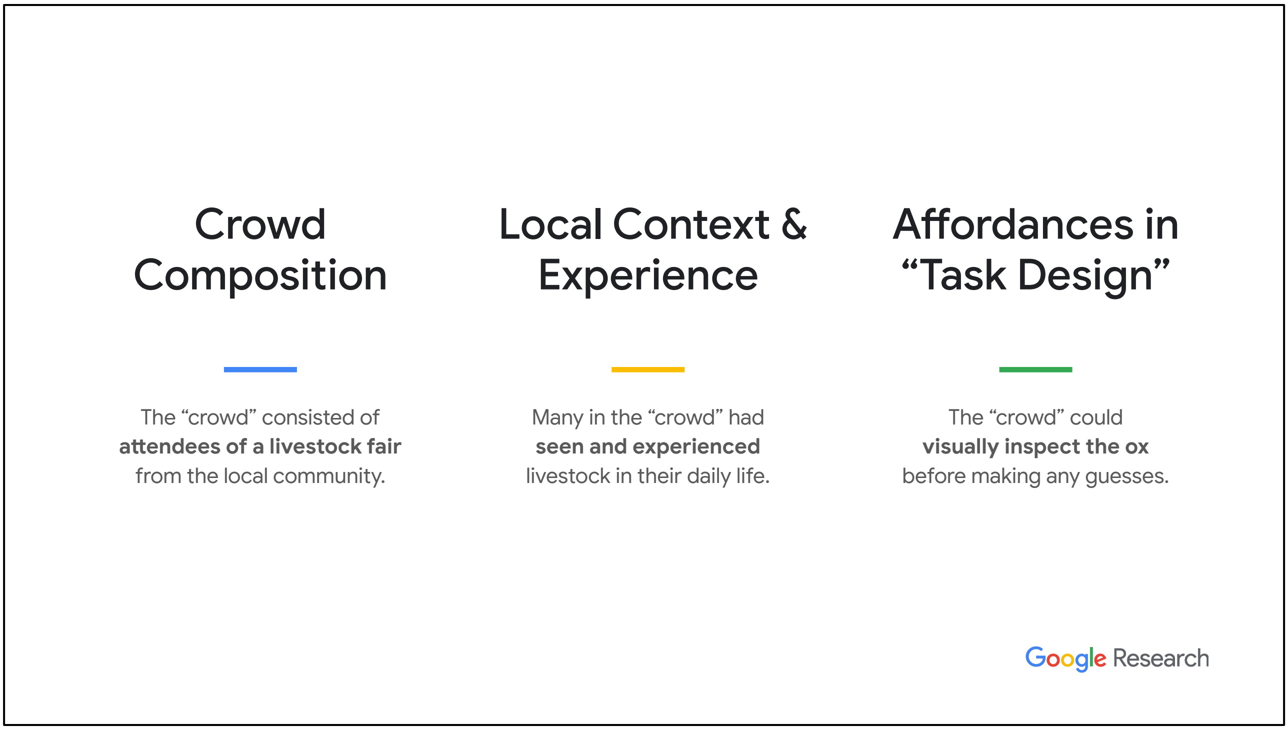
Google Research

People can be very optimistic about the ability of crowds in the aggregate. They often refer to the anecdote by Francis Galton of a contest at a livestock fair in 1906:

- Ox was on display, and ~800 participants were invited to guess the animal's weight after it was slaughtered and dressed
- Median of 1,207 pounds was within 0.8% of the weight measured by the judges.
- Mean of 1,197 pounds had almost zero error.

Image credits:

- Book cover is fair use: https://en.wikipedia.org/wiki/The_Wisdom_of_Crowds#/media/File:Wisecrowds.jpg
- Francis Galton photograph is public domain: https://en.wikipedia.org/wiki/Francis_Galton#/media/File:Francis_Galton_1850s.jpg
- Oxen drawing is CC BY 2.0: https://en.wikipedia.org/wiki/Ox#/media/File:Ploughing_with_Oxen.jpg



Crowd Composition

The "crowd" consisted of **attendees of a livestock fair** from the local community.

Local Context & Experience

Many in the "crowd" had **seen and experienced** livestock in their daily life.

Affordances in "Task Design"

The "crowd" could **visually inspect the ox** before making any guesses.

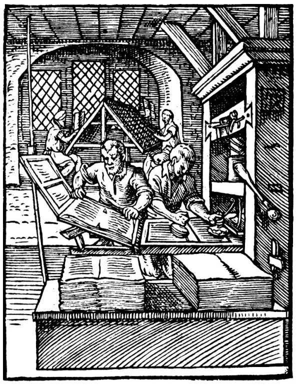
Google Research

The takeaway that "the wisdom of the crowd" always prevails is an overgeneralization. Some observations:

- The "crowd" was not just any group of people, but rather attendees of a livestock fair who likely had context and experience in the matter, even if they were not professional butchers
- The "task" was not a blind guess, but rather participants were allowed to visually inspect the animal to gather more cues

So the main challenge of crowdsourcing is finding the right "crowd" with the right knowledge to do a particular task, and then designing that task well enough to present them the right context to perform the task.

Crowdsourcing as a Design Problem



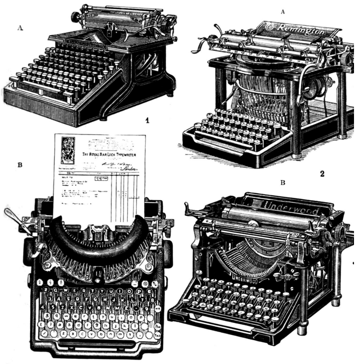
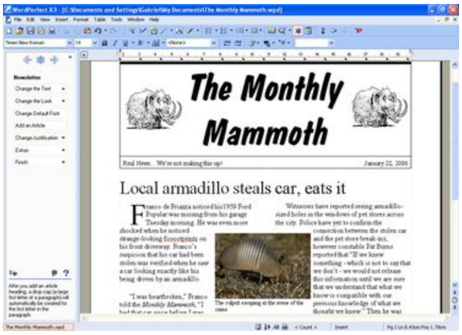


Fig. 1. Machines à un caractère par touche : A, Smith Premier ; B, Bar-Lock.

Fig. 2. Machines à 2 caractères par touche : A, Remingt B, Underwood.



After you add an article heading, a drop cap (a large first letter of a paragraph) will automatically be created for the first letter in the paragraph.

The Monthly Mammoth

Real News... We're not making this up! January 22, 2006

Local armadillo steals car, eats it

Franco de Friazza noticed his 1959 Ford Popular was missing from his garage Tuesday morning. He was even more shocked when he noticed strange-looking footprints on his front driveway. Franco's suspicion that his car had been stolen was verified when he saw a car looking exactly like his being driven by an armadillo.

"I was heartbroken," Franco told the *Monthly Mammoth*, "I had that car since before I was

The culprit escaping at the scene of the crime

Witnesses have reported seeing armadillo-sized holes in the windows of pet stores across the city. Police have yet to confirm the connection between the stolen car and the pet store break-ins, however constable Pat Burns reported that "If we knew something - which is not to say that we don't - we would not release this information until we are sure that we understand that what we know is compatible with our previous knowledge of what we thought we knew." Then he was

- Crowdsourcing is not only a data collection or tooling problem. Rather, it is similar to any other user-facing product opportunity: how do we design a context that best enables people to contribute their knowledge and judgment to the task at hand?

- Consider the development of typesetting:
  - Historically only experts could use the printing press to create publications
  - Then came typewriters, which allowed anyone to put type onto a page, but in a very restricted format
  - HCI/UX research led to the development of the word processor which allowed for near-expert results (depending on the user)

- We should think of crowdsourcing as a *design* problem: how can we use HCI/UX ideas (and understanding of the relevant domain for the problem) in order to motivate and enable people to contribute their crucial knowledge to data sets?

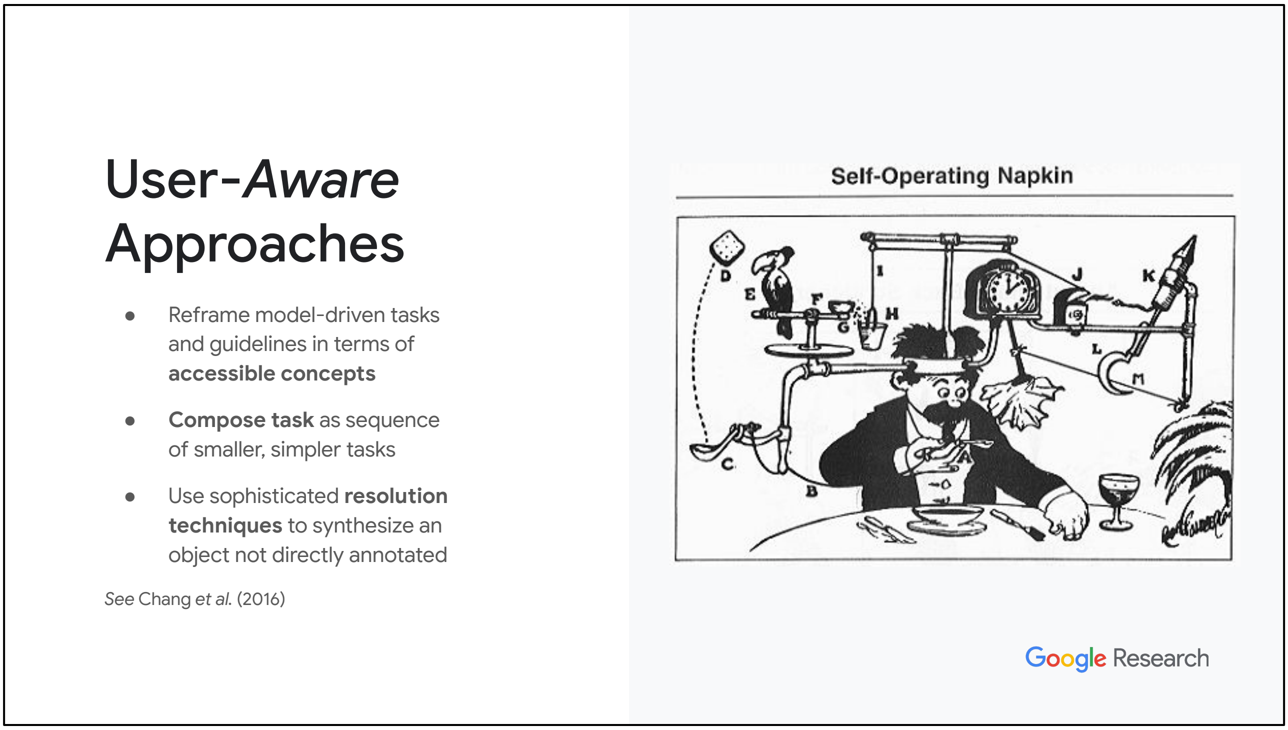
Image credits:

- Wooodcut is public domain: https://en.wikipedia.org/wiki/Printing_press#/media/File:Printer_in_1568-ce.png
- Typewriter diagram is public domain: https://en.wikipedia.org/wiki/Typewriter#/media/File:Comparison_of_Full-Keyb

- oard,_Single-Shift,_and_Double-Shift_Typerwriters_in_1911.png
- Screenshot is fair use: https://en.wikipedia.org/wiki/Word_processor#/media/File:WordPerfectX3.png

02

Linguistic Crowdsourcing



User-*Aware* Approaches

- Reframe model-driven tasks and guidelines in terms of **accessible concepts**
- **Compose task** as sequence of smaller, simpler tasks
- Use sophisticated **resolution techniques** to synthesize an object not directly annotated

See Chang *et al.* (2016)

Self-Operating Napkin

Google Research

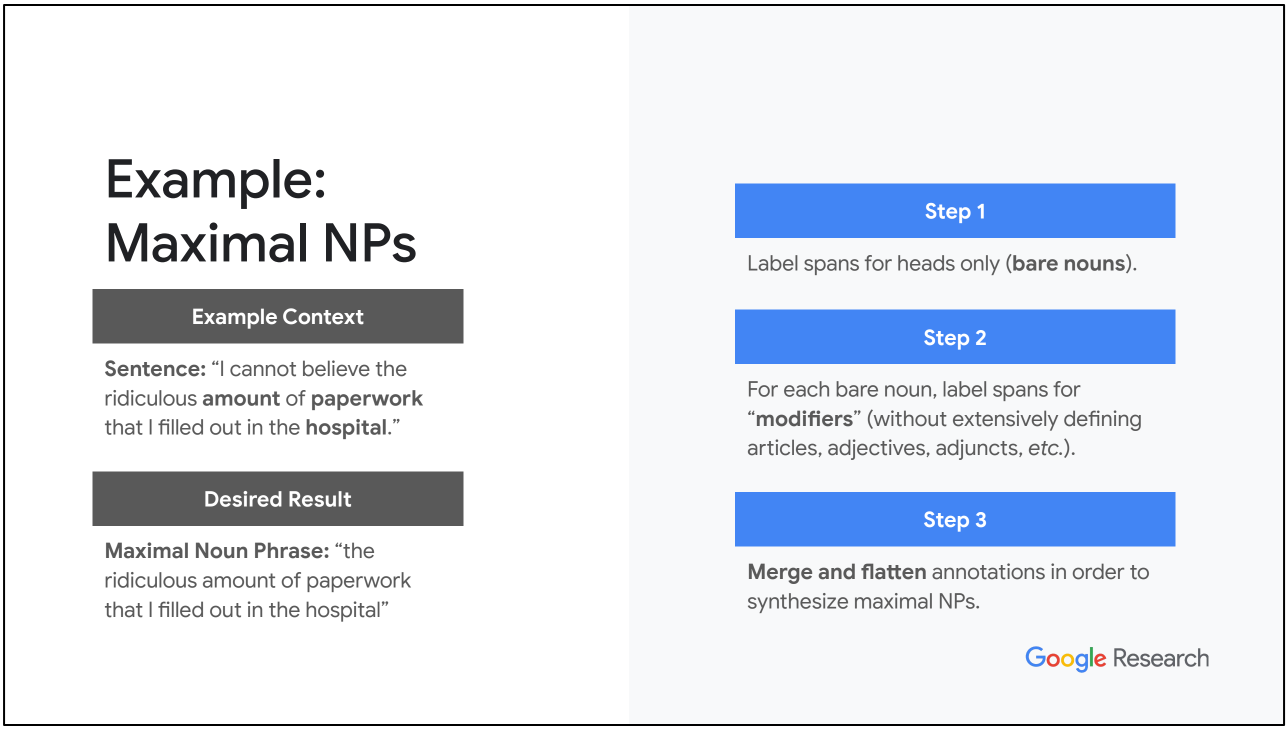
- In the case of linguistic annotation or linguistic data in general, many people have said that these kinds of judgments are too difficult for the average person. It would be too hard to find the right “crowd” so we should just rely on professional linguists.

- But the prospect of empowering “average” people, who are competent at their own language, to contribute to linguistic data sets has always been tantalizing.

- There is much research on ways we can “accommodate” or “compensate” for the types of “errors” people make or the differences in mental models that we can expect non-experts to have from experts. While this approach is designed with the user in mind, it does require a fairly elaborate apparatus of task sequencing, resolution, etc.

Image credits:

- Cartoon is public domain: https://en.wikipedia.org/wiki/Rube_Goldberg_machine#/media/File:Rube_Goldberg's_%22Self-Operating_Napkin%22_(cropped).gif



Example: Maximal NPs

Example Context

Sentence: “I cannot believe the ridiculous **amount** of **paperwork** that I filled out in the **hospital**.”

Desired Result

Maximal Noun Phrase: “the ridiculous amount of paperwork that I filled out in the hospital”

Step 1

Label spans for heads only (**bare nouns**).

Step 2

For each bare noun, label spans for “**modifiers**” (without extensively defining articles, adjectives, adjuncts, *etc.*).

Step 3

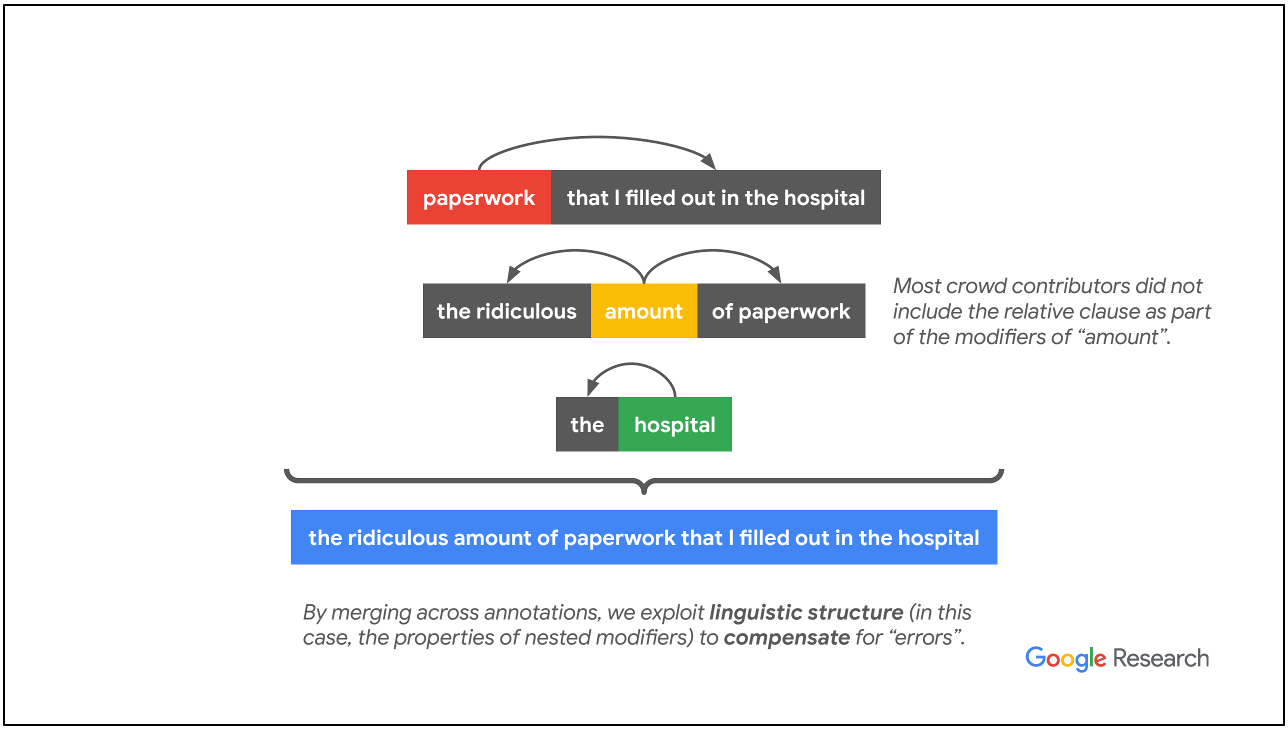
Merge and flatten annotations in order to synthesize maximal NPs.

Google Research

- As an example, suppose that we wanted to collect data on linguistic constituents (or dependency relationships) in order to augment data sets used to train and test a syntactic parsing model. The most straightforward approach would be to write some training guidelines and to ask people to label spans (within given sentences) corresponding to the structures of interest. We can rely on the knowledge that many people have an intuitive sense of certain linguistic structures (e.g., “a noun is a person, place, or thing”).

- But it turns out that even asking people to find something seemingly simple as a “noun phrase” is tricky. The “maximal projection” of a noun phrase can contain phrases or clauses that a non-linguist reader may not associate with the head noun. For example, if a higher-level noun phrase’s head noun is modified by a phrase that contains another noun phrase with its own phrasal modifiers, the non-linguist might not think that the nested modifiers have anything to do with the higher-level head noun.

- We broke down the task into (1) identification of bare nouns (fairly intuitive with some examples and diagnostics in the guidelines); and (2) identification of “modifiers” for each of the bare nouns. We did not ask questions about larger structures or ask people to annotate full noun phrases. Then we used a resolution method to merge the results from (2) into larger structures of interest.



paperwork that I filled out in the hospital

the ridiculous amount of paperwork

the hospital

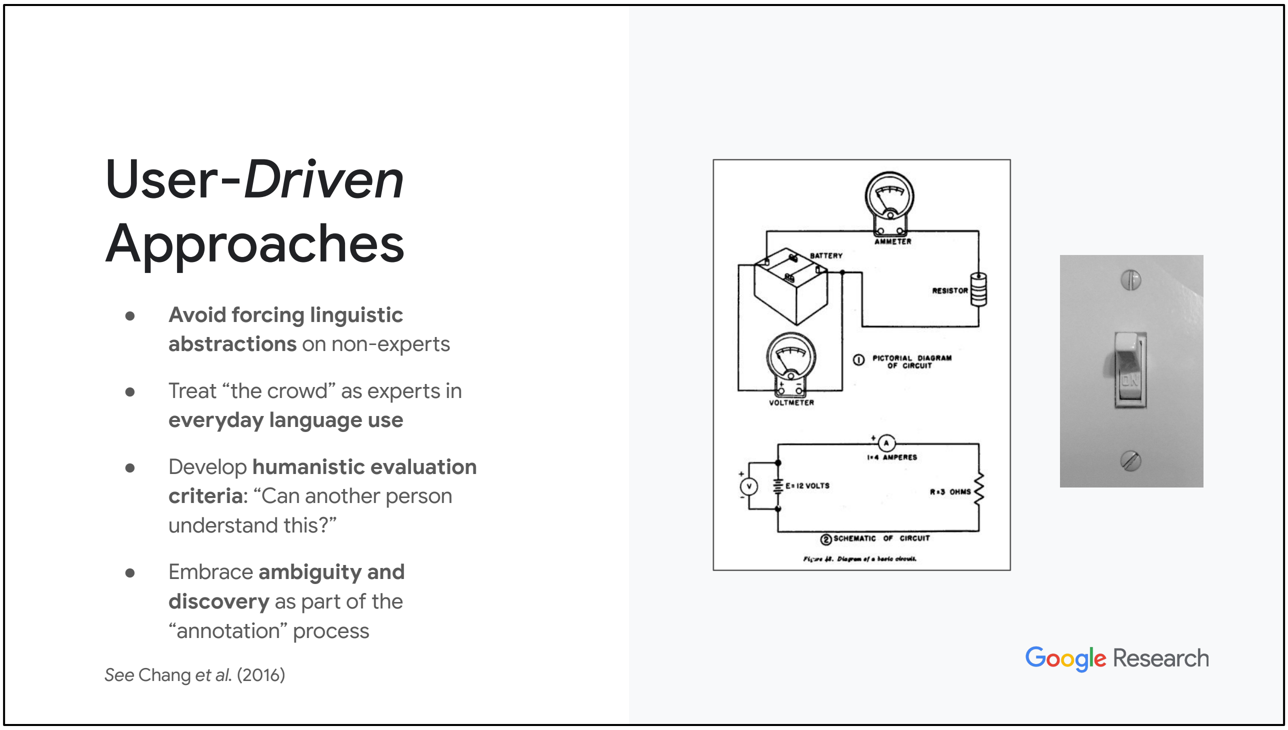
the ridiculous amount of paperwork that I filled out in the hospital

Most crowd contributors did not include the relative clause as part of the modifiers of “amount”.

By merging across annotations, we exploit ***linguistic structure*** *(in this case, the properties of nested modifiers) to* ***compensate*** *for “errors”.*

- Notably, the second annotation for “amount” is not a complete noun phrase in the context of the original syntax, but this “error” can be corrected when merging the results.

- This was more an exercise in feasibility — setting up a pipeline would likely be less practical than working with expert linguist annotators. It also would not necessarily work in other languages, or even within the same language for other structures of interest (e.g., verb phrases, non-contiguous structures).



User-*Driven* Approaches

- **Avoid forcing linguistic abstractions** on non-experts
- Treat “the crowd” as experts in **everyday language use**
- Develop **humanistic evaluation criteria**: “Can another person understand this?”
- Embrace **ambiguity and discovery** as part of the “annotation” process

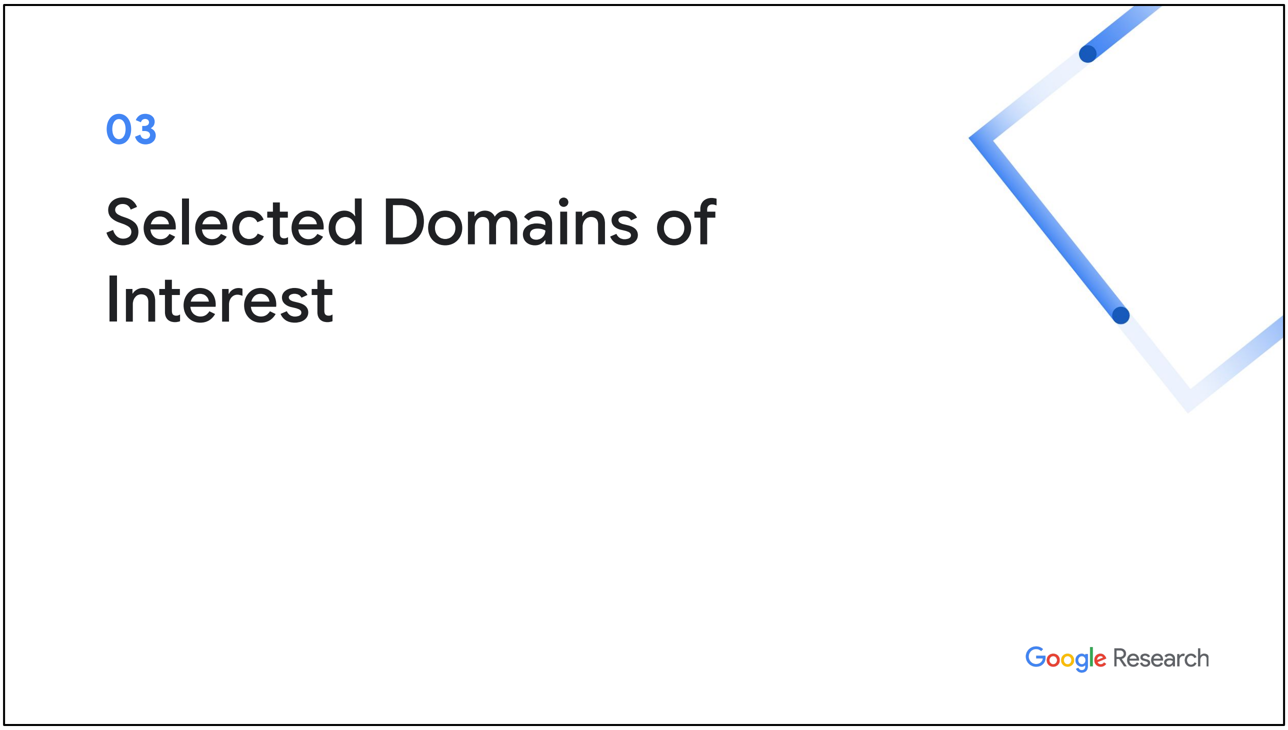
See Chang *et al.* (2016)

- Emphasis has shifted toward embracing the unconscious competence of non-experts, finding ways of asking questions that they understand intuitively, and allowing a variety of answers that still tell us something useful about the domain of interest. By analogy, most people do not know the internal workings of an electrical circuit, but they can reason about their experience with e.g. light switches

- In general, the paradigms of many language technologies are also moving in this direction (more end-to-end systems, and less emphasis on systems that analyze specific linguistic structures).

Image credits:

- Circuit diagram is public domain: https://en.wikipedia.org/wiki/Circuit_diagram#/media/File:Circuit_diagram_%E2%80%93_pictorial_and_schematic.png
- Light switch photograph is CC BY-SA 4.0: https://commons.wikimedia.org/wiki/File:A_double_toggle_light_switch.jpg#/media/File:A_double_toggle_light_switch.jpg



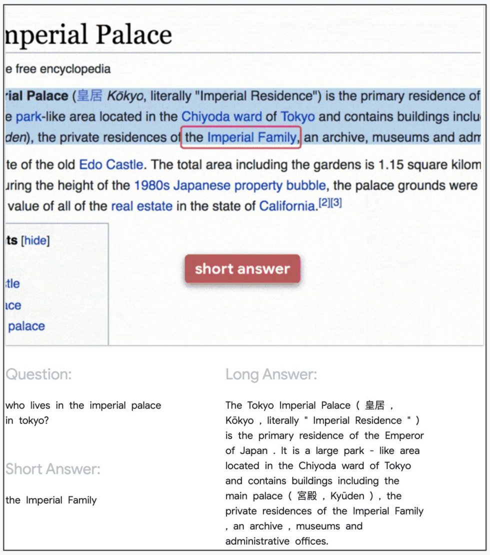
03

Selected Domains of Interest

Question Answering

- What **sections** of a document answer a given question? What are the **best-justified answers**?
- How do we collect **interesting** and **realistic** questions?
- How do we model and account for **ambiguity**?

See Kwiatkowski *et al.* (2019) and blog post; Clark *et al.* (2020) and blog post; Min *et al.* (2020)



nperial Palace

e free encyclopedia

ial Palace (皇居 *Kōkyo*, literally "Imperial Residence") is the primary residence of e park-like area located in the Chiyoda ward of Tokyo and contains buildings inclu *den*), the private residences of the Imperial Family, an archive, museums and adr te of the old Edo Castle. The total area including the gardens is 1.15 square kilom ıring the height of the 1980s Japanese property bubble, the palace grounds were value of all of the real estate in the state of California.[2][3]

Question:

who lives in the imperial palace in tokyo?

Short Answer:

the Imperial Family

Long Answer:

The Tokyo Imperial Palace (皇居 , Kōkyo , literally " Imperial Residence ") is the primary residence of the Emperor of Japan . It is a large park - like area located in the Chiyoda ward of Tokyo and contains buildings including the main palace (宮殿 , Kyūden) , the private residences of the Imperial Family , an archive , museums and administrative offices.

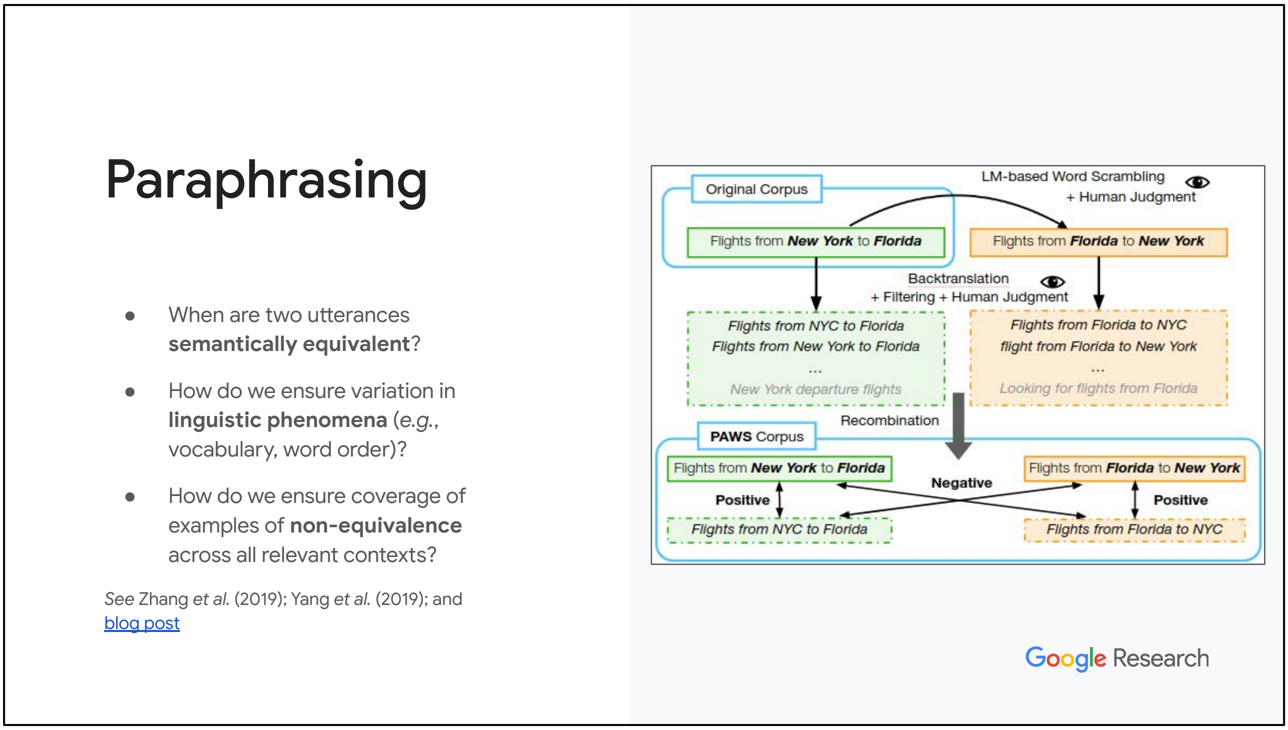
Google Research

- Many QA data sets derive the questions from documents, which can lead to highly selective coverage of questions (and a less interesting problem, if the answer is guaranteed to be in the document — real people often ask questions that aren't clearly answered in any one source).

- Compare to the Natural Questions and TyDiQA, which come from more organic methods, such as aggregating frequent questions that were issued to Google or even requesting that crowd contributors come up with new questions that are *not* answered by a prompt (as a seed for a second task, which is to find answers for this novel question on a given retrieved document). This is closer to the natural way that curious people explore possible answers questions raised by what they read.

Image credits:

- Screenshot: https://ai.google.com/research/NaturalQuestions/visualization



Paraphrasing

- When are two utterances **semantically equivalent**?
- How do we ensure variation in **linguistic phenomena** (*e.g.*, vocabulary, word order)?
- How do we ensure coverage of examples of **non-equivalence** across all relevant contexts?

See Zhang *et al.* (2019); Yang *et al.* (2019); and blog post

Original Corpus

LM-based Word Scrambling + Human Judgment

Flights from ***New York*** to ***Florida***

Flights from ***Florida*** to ***New York***

Backtranslation + Filtering + Human Judgment

Flights from NYC to Florida
Flights from New York to Florida
...
New York departure flights

Flights from Florida to NYC
flight from Florida to New York
...
Looking for flights from Florida

Recombination

PAWS Corpus

Flights from ***New York*** to ***Florida***

Flights from ***Florida*** to ***New York***

Negative

Positive

Positive

Flights from NYC to Florida

Flights from Florida to NYC

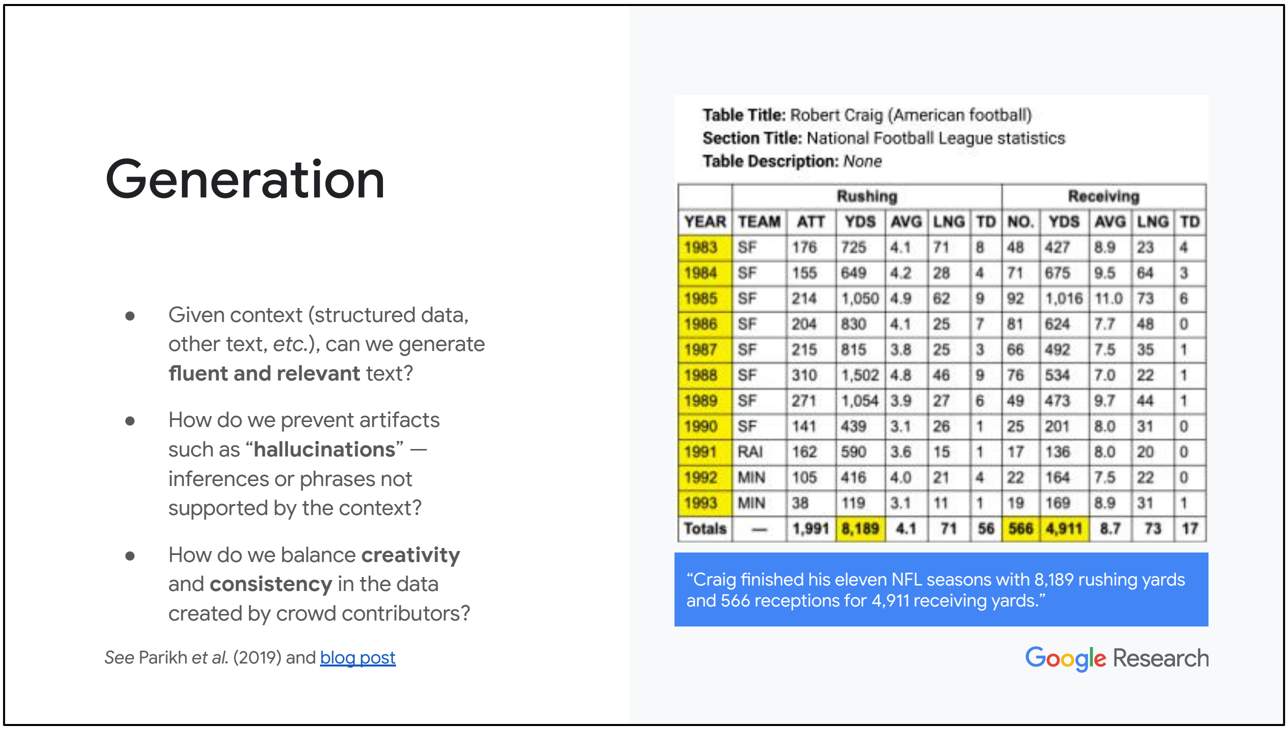
Google Research

- Many paraphrase data sets do not represent the full distribution of paraphrases that a more robust model would need for training and testing. They also do not contain negative examples (e.g., of critical non-paraphrase pairs such as “flights from New York to San Francisco” vs. “flights from San Francisco to New York”).

- Paraphrasing models would benefit from data sets with much higher variation, so that they do not penalize valid outputs that are very different from the training or testing data. That is, for each input sentence or paragraph, there should be a large variety of valid output paraphrases or summaries that exhibit many different linguistic variations, including word substitution/deletion, morphological inflection, word or phrase movement, clause restructuring, and so on.

- For PAWS, crowd contributors were not asked to write paraphrases from scratch — this likely would have been a daunting task and could have resulted in artifacts in the data set because of individual contributors’ writing styles. Instead, researchers used computational techniques (word scrambling, backtranslation) to ensure balanced coverage of certain features that they wanted in the data set. The human crowd contributors were then asked questions about the candidate paraphrase pairs. This enabled them to contribute nuanced semantic judgments and avoided the problem of asking them to pay attention to technical linguistic features

Image credits:
https://ai.googleblog.com/2019/10/releasing-paws-and-paws-x-two-new.html



Generation

- Given context (structured data, other text, *etc.*), can we generate **fluent and relevant** text?
- How do we prevent artifacts such as "**hallucinations**" — inferences or phrases not supported by the context?
- How do we balance **creativity** and **consistency** in the data created by crowd contributors?

See Parikh *et al.* (2019) and blog post

Table Title: Robert Craig (American football)
Section Title: National Football League statistics
Table Description: *None*

	Rushing						Receiving				
YEAR	TEAM	ATT	YDS	AVG	LNG	TD	NO.	YDS	AVG	LNG	TD
1983	SF	176	725	4.1	71	8	48	427	8.9	23	4
1984	SF	155	649	4.2	28	4	71	675	9.5	64	3
1985	SF	214	1,050	4.9	62	9	92	1,016	11.0	73	6
1986	SF	204	830	4.1	25	7	81	624	7.7	48	0
1987	SF	215	815	3.8	25	3	66	492	7.5	35	1
1988	SF	310	1,502	4.8	46	9	76	534	7.0	22	1
1989	SF	271	1,054	3.9	27	6	49	473	9.7	44	1
1990	SF	141	439	3.1	26	1	25	201	8.0	31	0
1991	RAI	162	590	3.6	15	1	17	136	8.0	20	0
1992	MIN	105	416	4.0	21	4	22	164	7.5	22	0
1993	MIN	38	119	3.1	11	1	19	169	8.9	31	1
Totals	**—**	**1,991**	**8,189**	**4.1**	**71**	**56**	**566**	**4,911**	**8.7**	**73**	**17**

"Craig finished his eleven NFL seasons with 8,189 rushing yards and 566 receptions for 4,911 receiving yards."

Google Research

- Data sets that rely only on naturally occurring sentences (from an existing corpus) may not be well-suited for a particular generation problem. Asking crowd contributors to generate sentences "from scratch" may result in data sets that exhibit both little creativity (because of the context in which the task is being performed) and also highly divergent writing styles, as well as the potential to introduce "hallucinated" facts not supported by the context.

- One approach to mitigate these issues is to control the annotation process much more closely — allowing the annotators to be creative within constraints. In the case of ToTTo, the task was transformed into a hybrid process where naturally occurring sentences related to a table (by a heuristic) were presented to annotators for further revision.

Image credits:

- Screenshot: https://ai.googleblog.com/2021/01/totto-controlled-table-to-text.html

Thank You

Q&A / Discussion

Google Research

Selected References

- Chang *et al*. 2016. Linguistic Wisdom from the Crowd. In *Crowdsourcing Breakthroughs for Language Technology Applications* (HCOMP 2015).
- Clark *et al*. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. In *TACL*.
- Kwiatkowski *et al*. 2019. Natural Questions: A Benchmark for Question Answering Research. In *TACL*.
- Min *et al*. 2020. AmbigQA: Answering Ambiguous Open-domain Questions. In *Proc. EMNLP*.
- Parikh *et al*. 2020. ToTTo: A Controlled Table-To-Text Generation Dataset. In *Proc. EMNLP*.
- Yang *et al*. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proc. EMNLP*.
- Zhang *et al*. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. NAACL*.

Google Research