CS388: Natural Language Processing

Lecture 9: Annotation + Dataset Bias

Greg Durrett

TEXAS
The University of Texas at Austin

Administrivia

- OPTIONAL LECTURE; normal lectures resume on Thursday
- Mini 2 due March 2
- +3 slip days
- Rest of the course pushed back

This Lecture

- Annotation practices + examples
- Datasheets for datasets
- Annotation artifacts, evolving datasets

Annotation

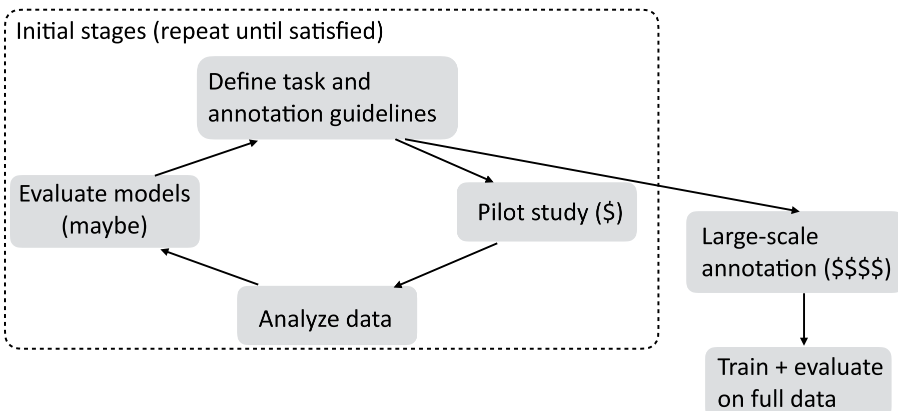
- A **critical** part of the ML pipeline
- Powerful models like neural networks (and BERT specifically) can learn patterns in the data — we need the right datasets to teach them the right patterns!
- **How do we build a good dataset?**

Annotation Practices

Annotating a Dataset

- Who's involved?
  - Researchers: you!
  - Annotators: typically people you hire, might be workers on platforms like Amazon Mechanical Turk
  - Stakeholders: whose data are you annotating / who will be impacted by the system?

Annotation Lifecycle

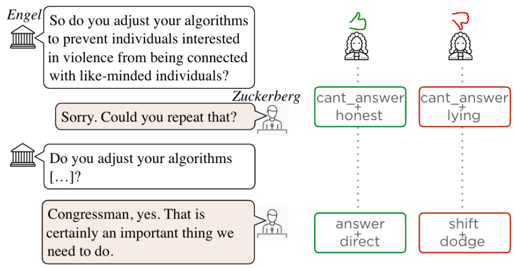


Defining the Task

- What is the goal of the annotation?
- How can you explain the task to annotators?
  - If using non-experts, how can linguistic tasks be communicated?

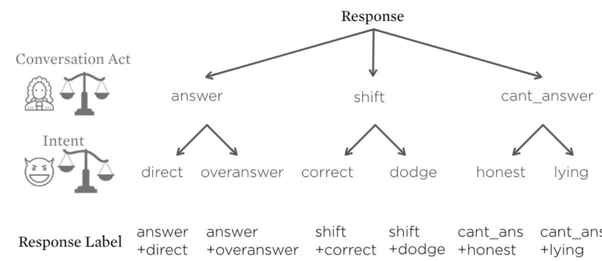
Example: Discourse Acts

- Annotate *perceived conversational intents* in Congressional hearings
- Annotators: workers on MTurk
- Stakeholders: researchers on discourse, social scientists
- Key focus: natural disagreement between the annotators based on their views of the speakers



Elisa Ferracane, Greg Durrett, Junyi Jessy Li, Katrin Erk. In submission

Example: Discourse Acts



- Other labels are possible (stalling), or more complex linguistic notions, but annotators then struggled to apply these consistently
- It is not easy to come up with the correct taxonomy here!

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, Katrin Erk. In submission

Defining the Task

- What is the goal of the annotation?
- How can you explain the task to annotators?
  - If using non-experts, how can linguistic tasks be communicated?
- How to make the task more engaging for annotators? Asking them to do something creative or a "challenge" is best!

Example: Regex Descriptions

Lines starting with a capital letter not containing the string "dog"

seq2seq model

```
concat(<cap>, .*) & ~contain("dog")
```

- Goal: collect pairs of (English description, regex code)
- Annotators: MTurk; Stakeholders: people who use regexes who need a system to generate them
- **How to get such pairs from non-programmers? How to ensure these pairs are realistic?**

Xi Ye, Qiaochu Chen, Isil Dillig, Greg Durrett. ACL 2020

Example: Regex Descriptions

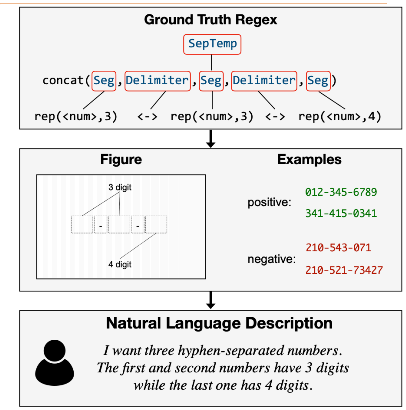
*The input will be in the form a **colon (:) separated tuple of three values**. The **first value** will be an integer (potentially a long in terms of size/length), with **the other two values** being either numeric or a string.*

- Realistic examples contain **referring expressions**, new **abstract constructs**
- How to get complex, realistic examples like this and not simple examples? If you ask people to write down a random regex task, they will come up with something simple
- We need to structure this task appropriately!

Xi Ye, Qiaochu Chen, Isil Dillig, Greg Durrett. ACL 2020

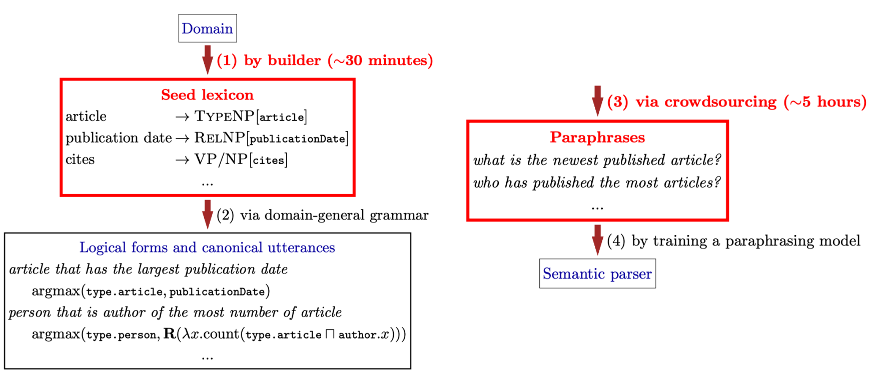
Example: Regex Descriptions

- Generate the ground-truth regex **first**, draw it as a figure, get people to describe it
- Annotators enjoyed this task (they emailed us!) and came up with creative descriptions



Xi Ye, Qiaochu Chen, Isil Dillig, Greg Durrett. ACL 2020

Data Collection "Overnight"



Yushi Wang et al. (2015)

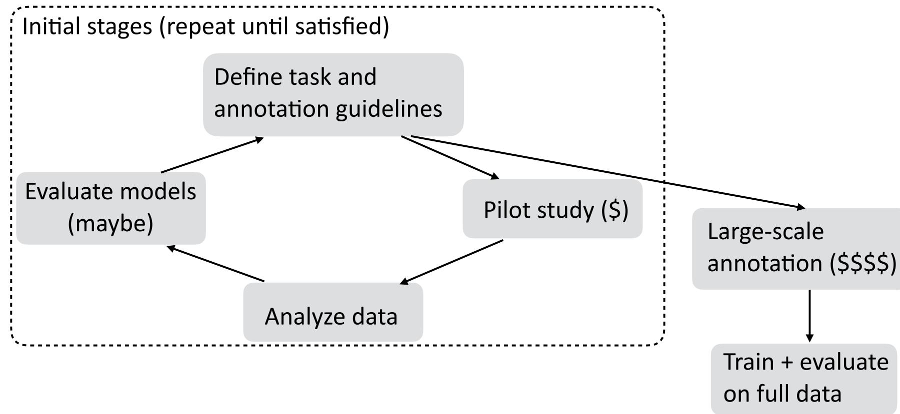
Pilot Studies

- Usually start with a small group of experts (e.g., the researchers and their colleagues/friends) and scale out to a group of non-experts
- Aim: collect enough data to assess annotator agreement and to tell what pitfalls might exist in the data

Analyzing Data

- How well do annotators agree?
- Metrics for categorical labels (e.g., multiclass problems): Krippendorf's alpha, Fleiss's kappa
  - 0-1 measures where 0 is the agreement of random chance
  - For the conversations: overall Krippendorf's alpha = 0.494 ("moderate")
  - Conversation act: 0.652. Intent: 0.376. Intents are more subjective, so we expect higher disagreement here!
- Metrics for real-valued ratings: Spearman's rho (corrects for different scales of different annotators)

Annotation Lifecycle



Datasheets for datasets

Datasheets for Datasets

- Framework for describing why a dataset was created, what's in it, how it was collected, etc.
- Motivation
  - **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
  - **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
  - **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

Gebru et al. (2018)

Datasheets for Datasets

- Composition
  - Questions about type of data, subsampling, nature of the labels
  - Train/dev/test splits
  - Noise/errors
  - Confidential/sensitive data, data about vulnerable subpopulations, identifiability
  - Dangerous/upsetting data

Gebru et al. (2018)

Datasheets for Datasets

- Collection process
  - How was the data acquired?
  - Who was involved in the process?
  - Was consent obtained to collect the data?
  - Was IRB approval obtained?

Gebru et al. (2018)

Datasheets for Datasets

- Preprocessing
- Uses
- Distribution
- Maintenance
- **Datasheets outline a good set of questions to consider when undertaking an annotation efort**

Gebru et al. (2018)

Annotation Artifacts

Natural Language Inference

- NLI, also called textual entailment: three class classification task over pairs of sentences
  - Entailment: premise *implies* hypothesis
  - Neutral: premise *is unrelated to* hypothesis
  - Contradiction: hypothesis *cannot be true* if premise is true

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

- Caveat: these sentences are understood to be about the same scenario. And the judgments are usually somewhat subjective

Natural Language Inference

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

Gururangan et al. (2018)

- Why is something entailed?
  - Hypernymy: *A woman is doing X -> A person is doing X*
  - Quantification: *Everybody is selling X -> Someone is selling X*
  - Commonsense: *A woman is selling bamboo sticks -> A woman wants to earn money*
  - Temporal: *A woman is selling X all day -> A woman is selling X at 2pm*

Natural Language Inference

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

Gururangan et al. (2018)

- Why is something contradicted? Actually this is pretty specific!
  - *A man is selling iced tea*: this could be true! Not a contradiction
  - Negation: *A woman is not selling bamboo sticks*: we have to assume it's the same woman, which we typically assume
  - Commonsense: *A woman is relaxing, doing nothing*
  - Quantification: *No woman is selling bamboo sticks*

Natural Language Inference

- How was the dataset annotated?

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. *Example: For the caption "Two dogs are running through a field." you could write "There are animals outdoors."*
- Write one alternate caption that **might be** a **true** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "Some puppies are running to catch a stick."*
- Write one alternate caption that is **definitely** a **false** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "The pets are sitting on a couch." This is different from the* maybe correct *category because it's impossible for the dogs to be both running and sitting.*

- Very clever protocol! But the open-endedness + the given examples lead annotators into certain patterns!

Bowman et al. (2015)

Natural Language Inference

Gururangan et al. (2018); Poliak et al. (2018)

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

- To create neutral sentences: annotators *add information*
- To create contradictions: annotators *add negation*
- Models can do very well *without looking at the premise*

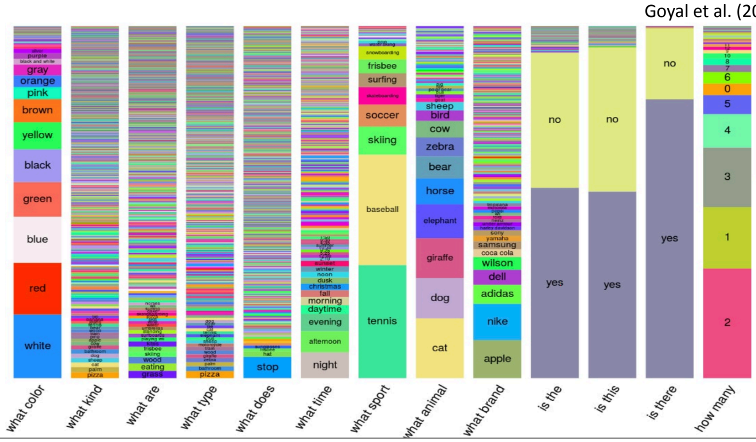
	Hypothesis-only	Majority	
SNLI	69.17	33.82	+35.35
MNLI-1	55.52	35.45	+20.07
MNLI-2	55.18	35.22	+19.96

What do we do?

- Why is this a problem? Because our models learn these simple cues and not actually the hard task we want them to learn
- They don't generalize to challenging new examples without these patterns — understanding this behavior is crucial to explaining what our models are doing!
- Solutions: build harder tasks, tweak data or training objective to inoculate models against this (many proposals)

Bias in Visual Question Answering

Goyal et al. (2018)



Visual Question Answering

- They collected multiple images with different answers for every question. Now the dataset is more balanced

Figure 1: Examples from our balanced VQA dataset.

Goyal et al. (2018)

Contrast Sets

- Construct *controlled* datasets that test what we want
- Perturb examples to highlight similar distinctions as in VQA

Original (Negative): I had quite high hopes for this film, even though it got a bad review in the paper. I was extremely tolerant, and sat through the entire film. I felt quite sick by the end.
New (Positive): I had quite high hopes for this film, even though it got a bad review in the paper. I was extremely amused, and sat through the entire film. I felt quite happy by the end.

Gardner et al. (2020)

Contrast Sets

Original (Positive): This is the greatest film I saw in 2002, whereas I'm used to mainstream movies. It is rich and makes a beautiful artistic act from these 11 short films. From the technical info (the chosen directors), I feared it would have an anti-American basis, but ... it's a kind of (11 times) personal tribute. The weakest point comes from Y. Chahine : he does not manage to "swallow his pride" and considers this event as a well-merited punishment ... It is really the weakest part of the movie, but this testifies of a real freedom of speech for the whole piece.
New (Negative): This is the most horrendous film I saw in 2002, whereas I'm used to mainstream movies. It is low budgeted and makes a less than beautiful artistic act from these 11 short films. From the technical info (the chosen directors), I feared it would have an anti-American basis, but ... it's a kind of (11 times) the same. One of the weakest point comes from Y. Chahine : he does not manage to "swallow his pride" and considers this event as a well-merited punishment ... It is not the weakest part of the movie, but this testifies of a real freedom of speech for the whole piece.

Gardner et al. (2020)

Dynamic Datasets

- Adversarial filtering (Le Bras et al., 2020): filter out data that is easily fit due to dataset biases
- Dynabench (FAIR): adaptive benchmarks with new data being collected to highlight errors
- Lots of ongoing work here!

Takeaways

- We looked at the basic procedures for constructing a dataset
- Lots of guiding frameworks, such as datasheets, for thinking about both data quality as well as possible ethical issues
- Dataset *biases*: these will come up again later!