Announcement

Assignment 1 is released and due on Feb 12, 11:59pm (ET).

Questions?

- Come to office hours or post them on Piazza.
- Important note: we will not answer assignment questions after the official due date (Feb 12).

CS 784: Computational Linguistics
Lecture 6: Datasets and Data Curation

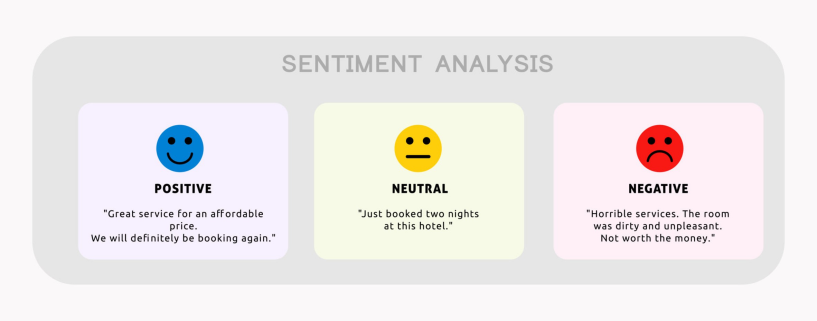
Freda Shi

School of Computer Science, University of Waterloo
fhs@uwaterloo.ca

January 23, 2025

Language Datasets with Computation

NLP datasets typically include **inputs** (usually text) and **outputs** (usually some sort of annotation).



Annotation

Supervised machine learning needs labeled datasets, where labels are called **ground truth**.

In NLP, most labels are annotations provided by humans.

There is always some disagreement among annotators, even for simple tasks.

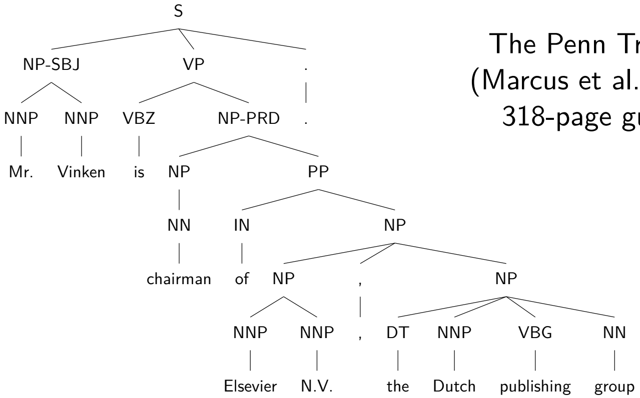
These annotations are called **gold standard**, not ground truth, although these terms are often used interchangeably.

When using labels generated by models for further training, we sometimes call them **silver standard**.

How are NLP/CL datasets developed?

Option 1 (traditional): paid & trained human annotators.

- Researchers write annotation guidelines, recruit & pay expert annotators.
- Consistent annotations but extremely costly to scale.



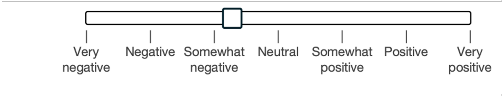
The Penn Treebank
(Marcus et al., 1993)
318-page guideline

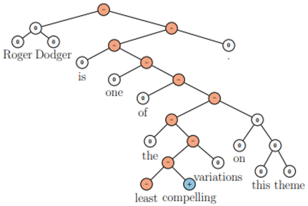
How are NLP/CL datasets developed?

Option 2 (modern): crowdsourcing.

- We can't really train annotators, but it's easier to get multiple annotations for each input (which can be averaged).

The Stanford Sentiment Treebank (SST; Socher et al., 2013)





Ethics in Crowdsourcing

A few questions to think about when conducting crowdsourcing:

- Will you exclude some participants based on some criteria?
- Will the participants interact with each other?
- Will the participants be paid? If so, how?
- Will you collect more data from the participants than you need?
- How will the data be stored?
- How will you share the research results with the participants?
- ...

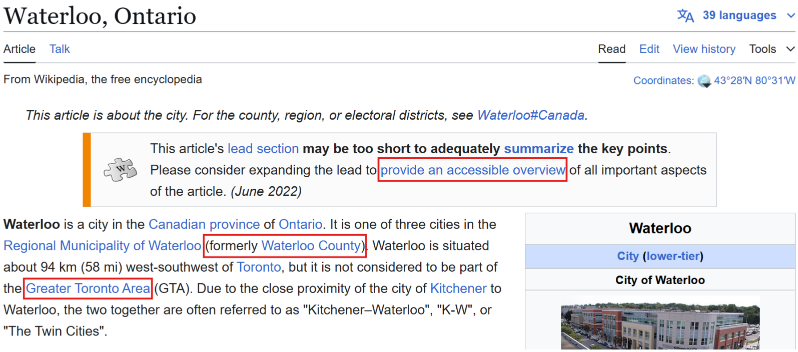
Consult this website before conducting experiments that involve human participants:
`https://uwaterloo.ca/research/office-research-ethics`

How are NLP/CL datasets developed?

Option 3 (modern): use *naturally occurring* annotations.

- Doesn't require any human annotation for the specific purpose.
- The data could be noisy, but it's often large-scale.

Any naturally-occurring annotations for parsing?



Waterloo, Ontario

39 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Coordinates: 43°28′N 80°31′W

This article is about the city. For the county, region, or electoral districts, see Waterloo#Canada.

This article's lead section **may be too short to adequately summarize the key points**. Please consider expanding the lead to provide an accessible overview of all important aspects of the article. *(June 2022)*

Waterloo is a city in the Canadian province of Ontario. It is one of three cities in the Regional Municipality of Waterloo (formerly Waterloo County). Waterloo is situated about 94 km (58 mi) west-southwest of Toronto, but it is not considered to be part of the Greater Toronto Area (GTA). Due to the close proximity of the city of Kitchener to Waterloo, the two together are often referred to as "Kitchener–Waterloo", "K-W", or "The Twin Cities".

Waterloo

City (lower-tier)

City of Waterloo

[Tianze Shi et al. NAACL 2021. Learning syntax from naturally-occurring bracketings.]

How are NLP/CL datasets developed?

In fact, naturally occurring annotations are the most common source of data nowadays.

We use web-text to pretrain language models!

There has been a trend towards using human-in-the-loop data collection, where humans are involved to provide feedback on the model's predictions.

Example: reinforcement learning with human feedback (RLHF; Ouyang et al., 2022).

Annotator Agreement: Agreement Percentage

Given annotations from two annotators, how should we measure the inter-annotator agreement?

- Agreement percentage

$$p_o = \frac{\sum_{i=1}^{n} \mathbb{1}[a_i = b_i]}{n}$$

n: number of examples $\mathbb{1}[\cdot]$: indicator function – 1 if the condition is true, 0 otherwise

Annotator Agreement: Cohen's Kappa

Given annotations from two annotators, how should we measure the inter-annotator agreement?

- Agreement percentage: $p_o = \frac{\sum_{i=1}^{n} \mathbb{1}[a_i = b_i]}{n}$
- Cohen's kappa

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

p_e: expected agreement by chance

A \ B	Y	N
Y	80	5
N	5	10

$$P_A(Y) = 0.85, P_A(N) = 0.15$$
$$P_B(Y) = 0.85, P_B(N) = 0.15$$

$$\begin{aligned} p_e &= P_A(Y)P_B(Y) + P_A(N)P_B(N) \\ &= 0.85 \times 0.85 + 0.15 \times 0.15 \\ &= 0.745 \end{aligned}$$

$$p_o = 0.9$$

$$\kappa = \frac{0.9 - 0.745}{1 - 0.745} = 0.608$$

Annotator Agreement: Cohen's Kappa (cont.)

Given annotations from two annotators, how should we measure the inter-annotator agreement?

- Agreement percentage: $p_o = \frac{\sum_{i=1}^{n} \mathbb{1}[a_i = b_i]}{n}$
- Cohen's kappa

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

p_e: expected agreement by chance

A \ B	Y	N
Y	45	5
N	5	45

$$\begin{aligned} P_A(Y) = 0.5, P_A(N) &= 0.5 \\ P_B(Y) = 0.5, P_B(N) &= 0.5 \end{aligned}$$

$$\begin{aligned} p_e &= P_A(Y)P_B(Y) + P_A(N)P_B(N) \\ &= 0.5 \times 0.5 + 0.5 \times 0.5 \\ &= 0.5 \\ p_o &= 0.9 \\ \kappa &= \frac{0.9 - 0.5}{1 - 0.5} = 0.8 \end{aligned}$$

Annotator Agreement: Fleiss' Kappa

Given annotations from two annotators, how should we measure the inter-annotator agreement?

- Agreement percentage: $p_o = \frac{\sum_{i=1}^{n} \mathbb{1}[a_i = b_i]}{n}$
- Cohen's kappa: $\kappa = \frac{p_o - p_e}{1 - p_e}$
- Fleiss' kappa: generalization of Cohen's kappa to more than 2 annotators and $c (c \geq 2)$ classes

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

$$\bar{P} = \frac{1}{n} \sum_{i=1}^{n} P_i \qquad P_i = \frac{1}{c(c-1)} \sum_{j=1}^{c} n_{ij}(n_{ij} - 1)$$

$$\bar{P}_e = \sum_{j=1}^{c} p_j^2 \qquad p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}$$

n_{ij} : # annotators who assigned item i to class j

n : # annotators $\qquad N$: # items

Be Careful with Dataset Curation

Measuring massive multitask language understanding (MMLU; Hendrycks et al., 2021) has become a popular benchmark in NLP, especially the development of large-scale language models.

Gema et al. Are We Done with MMLU? NAACL 2025
https://arxiv.org/abs/2406.04127

> *Maybe not. We identify and analyse errors in the popular Massive Multitask Language Understanding (MMLU) benchmark. Even though MMLU is widely adopted, our analysis demonstrates numerous ground truth errors that obscure the true capabilities of LLMs. For example, we find that* ***57% of the analysed questions in the Virology subset contain errors****. To address this issue, … we create MMLU-Redux, which is a subset of 5,700 manually re-annotated questions across all 57 MMLU subjects. We estimate that 6.49% of MMLU questions contain errors. Using MMLU-Redux, we demonstrate* ***significant discrepancies with the model performance metrics that were originally reported****…*

Next

Text Classification: Data, Features and Models