Dynamic Human Evaluation for Relative Model Comparison

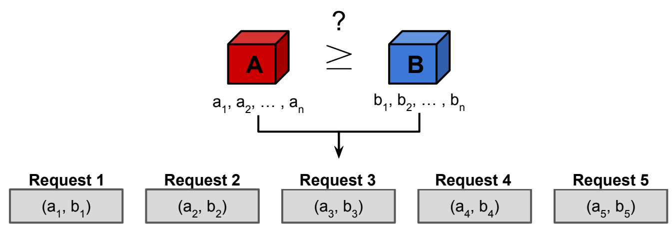
Thórhildur Thorleiksdóttir[1], Cedric Renggli[1], Nora Hollenstein[2], Ce Zhang[1]
[1]ETH Zürich; [2]University of Copenhagen

Evaluation of NLG Models

- Human evaluation is regarded as the primary metric
- Current limitations
  - Expensive and time consuming
  - Lack of consensus
  - Statistically underpowered

Model Comparison

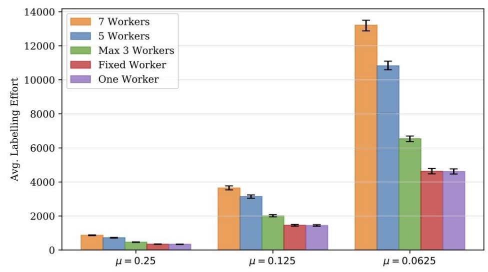
- Streamline human evaluation for text generation
- **Conclude better model with high probability**



- Two-alternative forced choice evaluation
- Control the number of collected judgements using Concentration Inequalities
- Compare different labelling strategies and their required labelling effort

Results

- Single random worker per request requires the **least labelling effort** when deciding the better model with **0.999** probability
- Assigning different workers per request enables trivial parallelization



- The human evaluation study indicated that assigning one random worker per request requires the least labelling effort in both model comparisons with a high probability (0.9999)
- Simulated and real human evaluation show similar trends in terms of labelling efforts for proposed decision method
- Simulating human evaluation can provide valuable insight without any cost

Agent-Based Human Evaluation

Simulate Two-Choice Human Evaluation

- Assume two generative models: **A** and **B**
- Varying workers evaluate provided request pairs → (a_i, b_i)
- Model performance: Proportion of selected outputs w.r.t. the number of requests evaluated

Formulation of the Evaluation Task

- **Request difficulty** $d \sim N(\mu, \sigma^2)$
  - **d = 1**, Easy to distinguish **a** as the better item compared to b
  - **d = 0**, Cannot distinguish a being better than b (and vice versa)
  - **d = -1**, Easy to distinguish **b** as the better item compared to a
- **Worker capacity** $c \sim \mathrm{Unif}(a, b)$
  - **c = 0**, Incapable annotator, not fluent in English
  - **c = 1**, Highly capable annotator, fluent in English
- **Compute the product to simulate the item selection**

$$p = c \cdot d$$

- **Transform to probability**

$$P(a) = \frac{p+1}{2} \qquad P(b) = 1 - P(a)$$

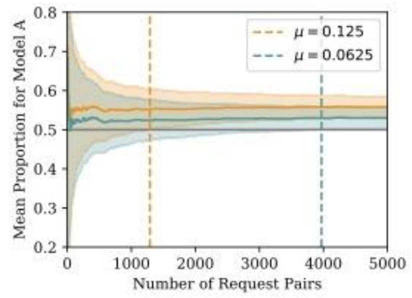
- **Perform a single Bernoulli Trial**

$$P(1) = P(a) \qquad P(0) = P(b)$$

Decision Boundaries

- **One-sided version of Hoeffding inequality** $\delta \leq e^{-2nt^2}$
  - **δ:** probability of the observed proportion not being within the error bounds
  - **t:** the width of the error bound
  - **n:** number of requests

$$t = \sqrt{\frac{-\ln(\delta)}{2n}}$$



Labelling Strategies

- Fixed Worker
- One Worker
- N Workers (Majority Vote)
- Max Three Workers

Experiment setup

- Simulation experiment consists of 1000 iteration for all labelling strategies where identical requests are evaluated with varying worker capabilities
- Sample 100 capabilities from $\mathrm{Unif}(0.8, 1.0)$
- Run simulation experiments with three different difficulty levels

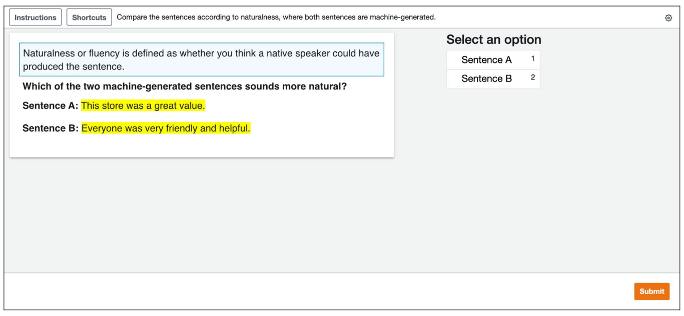
Case Study: Evaluating Controlled Text Generation

- Systematic control for semantic and syntactic aspects of generated text
- Train several versions of attribute-control text generation models
- **Two model comparisons:** V1 vs CGA and V2 vs CGA

Model	WD	Dataset Size
L_{ADV} + standard WD (V1)	0.3	$\sim$ 1300 sent.
L_{ADV} + standard WD (V2)	0.7	$\sim$ 600.000 sent.
L_{CGA} + cyclical WD (CGA)	ζ	$\sim$ 600.000 sent.

Experiment setup

- 500 request pair for each model comparison
- Evaluation Criteria: **Naturalness**
  Could a native speaker have produced the given text
- 10 workers evaluate each request pair on Amazon Mechanical Turk
- Sample collected judgments over 100 iterations



LREC 2022 Marseille

ETH zürich

UNIVERSITY OF COPENHAGEN