**Barch: an English Dataset of Bar Chart Summaries**

Iza Škrjanec, Muhammad Salman Edhi and Vera Demberg

**LREC Marseille, 20-25 June 2022**

---

**Introduction**

- Existing chart-summary datasets are either based on templates (Zhu et al. 2021) or include noisy summaries without additional alignment (Obeid and Hoque 2020)
- Basic data table entities vs analytical observations in chart summaries
- Research questions
  - What kind of analytical inferences do humans make in the summaries?
  - Do summaries differ when manipulating the main message (conveyed in the chart title)?

---

**Corpus creation**

1. **Chart design**: vertical bar charts from fabricated data
   - neutral: Average prices of digital cameras per brand
   - proportional: The most expensive digital cameras by average price
   - inverse: The most affordable digital cameras by average price
   - emphasis: Average price of cameras by Memoto and other brands

2. **Summary crowdsourcing**
   - 72 native speakers of English recruited via Prolific, presented with a bar chart and instructed to describe it in a summary that should suffice for a good understanding of the data.

3. **Chart-summary alignment**
   - Semi-automatic alignment for basic and analytical chart entities in summaries

---

**Properties of human-written summaries**

<table>
<thead>
<tr>
<th>Analytical category</th>
<th>Count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height approximation</td>
<td>1,294 (74.93)</td>
</tr>
<tr>
<td>Group name</td>
<td>143 (8.28)</td>
</tr>
<tr>
<td>Group height</td>
<td>91 (5.27)</td>
</tr>
<tr>
<td>Multiplication</td>
<td>85 (4.92)</td>
</tr>
<tr>
<td>Addition</td>
<td>73 (4.23)</td>
</tr>
<tr>
<td>Slope</td>
<td>29 (1.68)</td>
</tr>
</tbody>
</table>

- **Most summaries include at least one analytical entity**:
  - height approx. most frequent
  - group references for bars with similar height
  - arithmetic operations with small integers or multiples of 5

- **Dominant narrative for neutral and proportional**: descending order starting with the highest bar.
- For inverse, the summaries mostly start with the lowest bar.
- In 71% of emphasis cases, summaries start with the focus bar.

---

**Humans produce analytical observations about the charts and order the entities in the summary according to the main message.**

---

**Conclusions**

1. Barch is a dataset suitable for studying how humans describe charts given different messages and topics. It can be used to train data-driven chart-to-text NLG models.
2. The chart title affects the entity order in human-written summaries.
3. Humans naturally draw analytical observations about chart data and produce them in text.
4. NLG: generated summaries do not show these properties. On top of that, the models struggle with either fluency or input fidelity when trained on a small dataset like Barch.
5. Future work: further explore pretraining and pool the existing chart-summary corpora to counter overfitting and hallucinations.

---

**Corpus statistics**

- 47 charts - 1,063 summaries
- 1 chart - 22 summaries
- Charts drawn from 18 topics (education, stock market, weather, gender equality, among others)
- 1 summary:
  - on average, 3 sentences
  - roughly 54 tokens

---

**Dataset**

https://github.com/izaskr/barch_dataset

---

**Properties of human-written summaries**

- Most summaries include at least one analytical entity:
  - height approx. most frequent
  - group references for bars with similar height
  - arithmetic operations with small integers or multiples of 5

---

**NLG experiments with neural seq2seq**

**Test domain**: un/seen at training
**Model input**: basic (+ analytical) key-value pairs

3 **models**: with (KGPT, Chen et al. 2020) and without pretraining (C2T, Obeid and Hoque 2020; LSTM with attention)

**Automatic (BLEU, GPT-2 perplexity) and human (grammaticality, support) evaluation**