Barch: an English Dataset of Bar Chart Summaries

Iza Škrjanec, Muhammad Salman Edhi and Vera Demberg

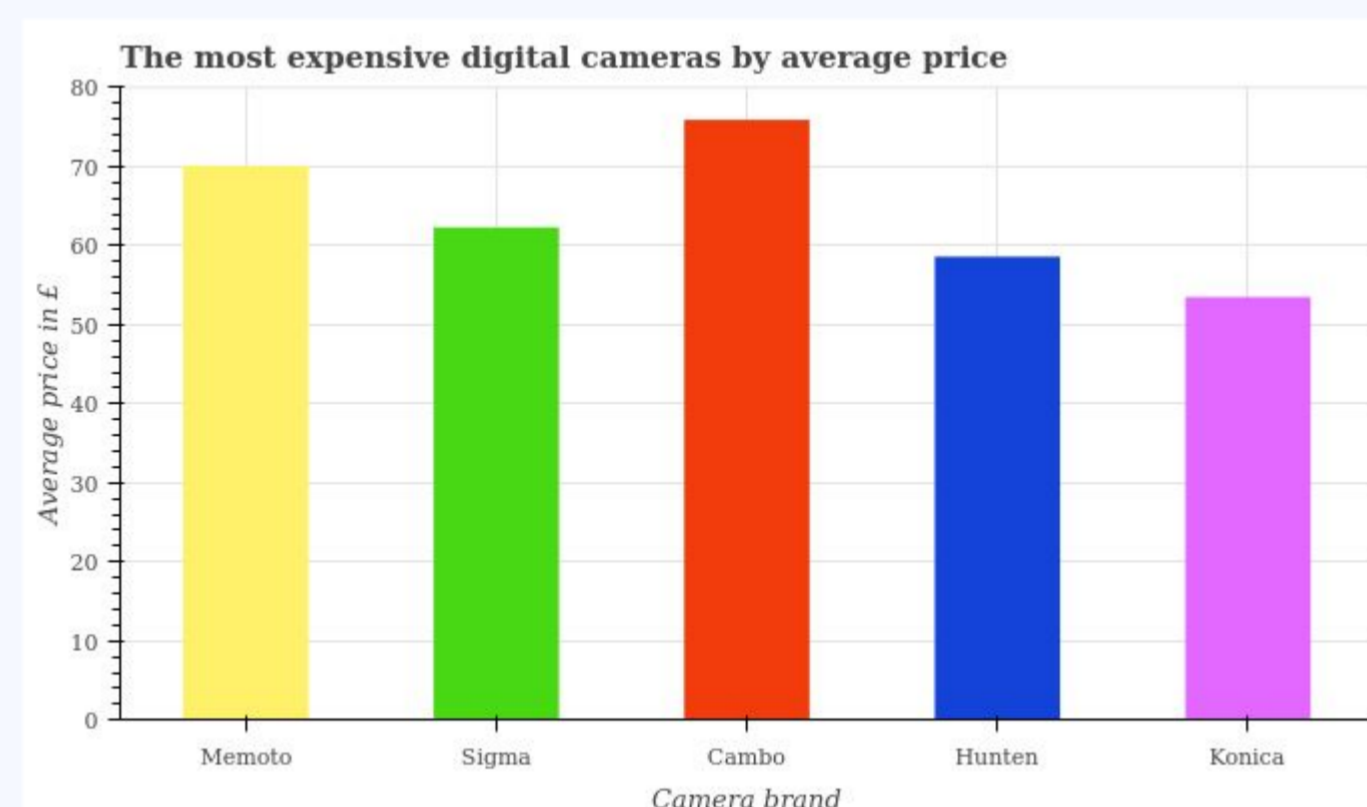*LREC* Marseille, 20-25 June 2022

UNIVERSITÄT DES SAARLANDES

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
The chart title conveys one of **4 intended messages**:
- 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
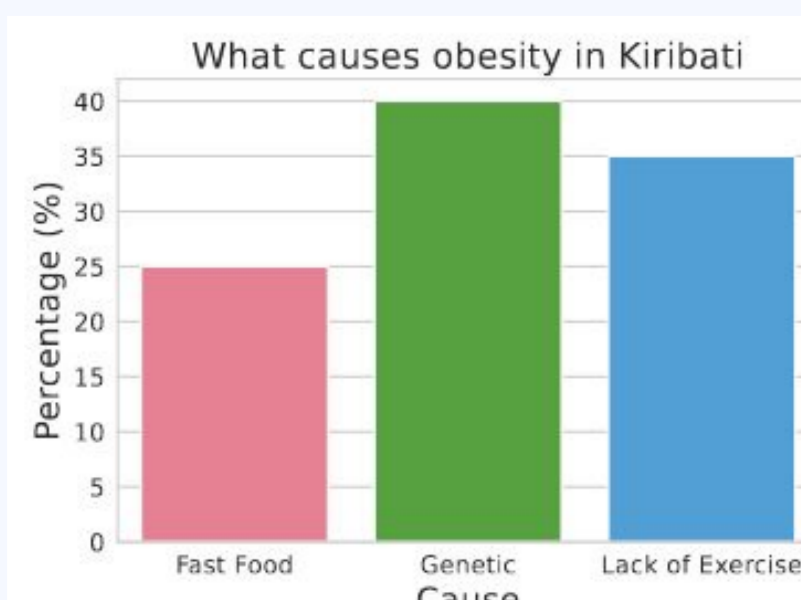- emphasis: Average price of cameras by Memoto and other brands

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

Analytical category	Count (%)
Height approximation	1,294 (74.93)
Group name	143 (8.28)
Group height	91 (5.27)
Multiplication	85 (4.92)
Addition	73 (4.23)
Slope	29 (1.68)

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

Message	% summaries starting with bar		
	highest	lowest	other
neutral	65.13	16.63	18.23
proportional	64.86	24.32	10.81
inverse	39.01	51.65	9.34

- 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.

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)

Test domain	Input data	Model	BLEU	GPT-2 ppl.	grammaticality	support (%)
seen	basic	LSTM	6.42	65.36	3.75	25.00
		C2T	20.85	66.06	4.30	80.00
		KGPT	**32.96**	18.71	**6.55**	37.21
	basic +analytical	LSTM	8.19	66.92	3.45	24.49
		C2T	26.16	41.05	4.90	**94.74**
		KGPT	19.90	**7.27**	5.10	38.89
unseen	basic	LSTM	19.84	62.44	3.17	7.14
		C2T	24.04	75.55	2.25	50.00
		KGPT	21.11	16.32	6.30	19.05
	basic +analytical	LSTM	20.57	123.98	3.08	11.76
		C2T	30.03	28.48	2.70	80.77
		KGPT	22.93	12.40	4.17	14.29

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

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
- Summaries per message:
  - neutral: 581
  - proportional: 184
  - inverse: 183
  - emphasis: 115

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.

Dataset

https://github.com/izaskr/barch_dataset