Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3552–3560
Marseille, 20-25 June 2022

Barch: an English Dataset of Bar Chart Summaries

Iza Škrjanec, Muhammad Salman Edhi, Vera Demberg
Saarland University
Saarbrücken, Saarland, 66123, Germany
{skrjanec, vera}@coli.uni-saarland.de, salman.edhi@gmail.com

Abstract

We present Barch, a new English dataset of human-written summaries describing bar charts. This dataset contains 47 charts based on a selection of 18 topics. Each chart is associated with one of the four intended messages expressed in the chart title. Using crowdsourcing, we collected around 20 summaries per chart, or one thousand in total. The text of the summaries is aligned with the chart data as well as with analytical inferences about the data drawn by humans. Our datasets is one of the first to explore the effect of intended messages on the data descriptions in chart summaries. Additionally, it lends itself well to the task of training data-driven systems for chart-to-text generation. We provide results on the performance of state-of-the-art neural generation models trained on this dataset and discuss the strengths and shortcomings of different models.

Keywords: chart summary, crowdsourcing, natural language generation

1. Introduction

Complex quantitative data is often visually presented in a chart and textually described in a summary in natural language. The general communicative goal of the summary is to inform the readers about the content of the chart and help them draw inferences about the data. When writing a summary of a chart such as that in (a) of Table 1 several choices have to be made in terms of the order and content: whether to state the height of each bar or of a selection, describe a general trend or include numerical inferences about the data and so on. Table 1 shows a bar chart, its underlying data table and a summary, the latter providing *basic information*: the height of each bar. The final sentence provides an inference comparing two bars, stating that one bar is nearly twice as high as the other one. This kind of information is *analytical*: it puts multiple data points in a relation (arithmetic operation) instead of simply describing single data points. Our data collection shows that human speakers frequently include such statements into chart summaries. However, many of existing datasets with chart and summaries do not consider such analytical information.

In fact, recent years have shown a growing interest in developing datasets with pairs of charts and their summaries with two prominent applications. On the one hand, chart-summary datasets can be used to study the production of summaries in humans, and can provide a data basis for studies which compare different descriptions in terms of how well they meet users' information needs.

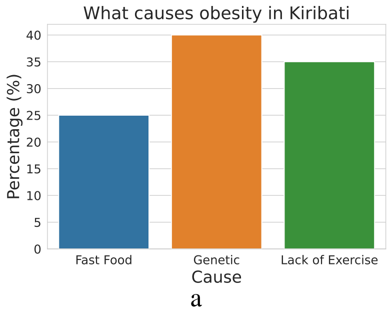
The chart design as well as the summary content can lead to variance in what inferences a comprehender draws from it. For example, Talbot et al. (2014) find that even the slightest changes in the chart design can result in a different perception and inferences by the user. It is key that the designers of the chart create the figure according to the message they intend to convey to the audience.

On the other hand, chart-summary datasets can serve as training sets for data-driven systems for natural language generation (NLG). Automating the task of writing chart summaries has a large potential in the context of data analytics and exploration. It also increases accessibility for users with vision impairments. Generating summaries from chart data falls under the task of data-to-text NLG, which has seen an impressive progress in output fluency and semantic fidelity (Clive et al., 2021; Chen et al., 2020; Li et al., 2021).

Over the last few years, two sizeable chart-summary datasets have been created. The dataset by Obeid and Hoque (2020) was crawled from online sources with human-written summaries of charts. It considerably differs in its data acquisition method from the dataset by Zhu et al. (2021), which contains synthetic data combined with automatically generated summaries based on templates. While template-based summaries have questionable ecological validity, the ones collected from online resources turn out to often include external information not limited on the chart data. The summaries also differ in the patterns of how they describe the chart data. The summaries by Zhu et al. (2021) contain analytical information beyond basic descriptions of heights and entity names, as they intentionally design their templates to state the general trends and other inferred information about the chart. In contrast, the chart-summary alignments in Obeid and Hoque (2020) cover only basic information.

In this paper, we address the need for human-written chart summaries with task-relevant analytical properties. We present Barch, an English dataset that lends itself to the analysis of human behavior as well as training of NLG models. The dataset includes:

- pairs of charts and human-written summaries (collected via crowdsourcing),
- charts with the same underlying data, but a different intended message signalled in the chart title,



Bar	**Height**
Fast Food	25
Genetic	40
Lack of Exercise	35
Title	What causes obesity in Kiribati
X-axis label	Cause
Y-axis label	Percentage (%)

b

This chart looked at causes of obesity in Kiribati. 25% was attributed to fast food and 35% to lack of exercise. The highest cause was genetic at 40% which was nearly twice as attributable to fast food.

c

Table 1: An example from our corpus. A bar chart (a), the data table with basic information (b), and an example of a crowdsourced summary that includes basic as well as analytical information about the chart.

- multiple summaries per chart,
- labels of semantic alignment between charts and summaries for basic and analytical information (missing in related datasets).

In Section 2 we review related work on chart summary datasets created for NLG. Section 3 describes the design and collection of Barch, followed by an analysis in Section 4. We trained neural NLG systems with one baseline and two state-of-the-art architectures and evaluated them with automatic as well as human evaluation (Sections 5 and 6). We conclude the paper and suggest further work in Section 7.

2. Related Work

While several datasets with chart-summary pairs have been created over the years, they either do not provide the material to study how humans naturally produce summaries under chart design manipulation or they lack some alignment between the chart and the summary, which can limit the output diversity of NLG models.

The Chart2Text dataset (Obeid and Hoque, 2020) is a large-scale collection of charts and summaries taken from Statista, an online platform of real-world charts and summaries collected from different sources and domains. Chart2Text includes over 8,000 bar and line plots with one human-written summary per chart.

The bar, line and scatter plots in AutoChart (Zhu et al., 2021) are based on real statistical data from different sources, but the authors designed the charts themselves. They generated summaries automatically using templates. The templates are designed to give a basic description of the data points and some analytical information about the trends or comparisons. For each of over 10,000 charts, two or more summaries were created.

Even though Chart2Text (Obeid and Hoque, 2020) contains human-written summaries, it cannot be used to explore the diversity in summaries of the same chart or whether changes in the chart design result in different summaries. On the other hand, AutoChart (Zhu et al., 2021) provides multiple summaries per chart, but the summaries and the differences between them are not ecologically valid.

There exist other relevant but publicly unavailable datasets, such as those based on templates (Ferres et al., 2007; Demir et al., 2012) or collected with crowdsourcing (Greenbacker et al., 2011).

A data-driven chart-to-text NLG model learns to map the chart data to its summary and this requires a suitable alignment between the source and the target side. The alignment is crucial in the substitution step when the chart values in the summary are substituted by placeholders. The Chart2Text (Obeid and Hoque, 2020) summaries often contain noise in the form of text that is not grounded in the input chart, but rather in external knowledge. Further, their substitution algorithm covers basic chart information, while analytical inferences are ignored, although they appear in the summaries. The AutoChart (Zhu et al., 2021) dataset does not have this issue because the summaries are strictly based on the charts and they include analytical information.

Despite the extensive data collection and creation of chart-summary pairs, there is missing data of human-written summaries suitable for analyzing what humans perceive as relevant in a chart, what information is included into a summary and in what order. A more detailed account of *basic* versus *analytical* chart information is also required. Training data-driven NLG system on such corpora is also promising in terms of output adequacy and diversity.

3. New Dataset for Chart-to-Summary Generation

To collect summaries, we first generate charts. The chart data is fabricated and drawn from different domains (topics). In addition, we manually specify four alternative intended messages or perspectives on a chart, which are conveyed by the chart title, see Table 2. The idea behind these four messages is that they will allow us to study how different intended messages shape chart summaries. After collecting the summaries via crowdsourcing, we align and annotate the text with the chart data. The corpus and annotation guidelines are available here: `https://github.com/izaskr/barch_dataset`.

3.1. Chart Design

We designed vertical bar charts with labeled axes and 3 to 6 bars. Each chart has a title that conveys one of the four intended messages: neutral, proportional, inverse or emphasis. These messages emphasise different bars of the chart: the “proportional” message indirectly emphasises the highest bar(s), the “inverse” the lowest bar(s), and the “emphasis” a specific bar of arbitrary height. Table 2 gives an example of titles for a chart with prices of cameras (y-axis) by different brands (x-axis). The chart with the proportional message is in Figure 1. In total, we designed 47 charts from 18 topics (more in Section 3.5).

Message	Title
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

Table 2: Examples of titles for each intended message for the same underlying chart data.

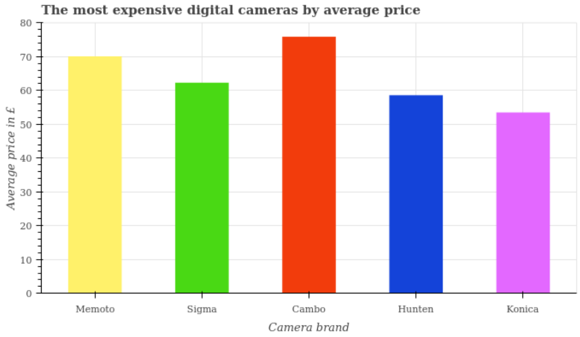


Figure 1: A bar chart with a proportional message conveyed in the title.

3.2. Data Collection

We crowdsourced the summaries online by recruiting the participants via Prolific[1] and collecting the data with LingoTurk (Pusse et al., 2016). In the experiment, participants were presented with a chart and asked to describe it as they would verbally present it to an audience. The summary should suffice for a good understanding of the data even in the case when the chart itself is not provided.

Before the start of the experiment, the instructions explained that the data in the charts is fabricated. An example of a chart and its summary was provided as well. The hourly pay rate of this experiment was 8.48 GBP. Native speakers of English from the United Kingdom with at least a Bachelor’s degree and no reading disabilities were allowed to take part. A total of 72 participants took part in the crowdsourcing process.

Upon collection, 16 summaries were discarded due to poor performance of the participants (nonfactuality or evident lack of effort), leaving about 22 summaries per chart on average or 1,063 summaries in total.

3.3. Annotation Categories

We define a set of **basic labels** to annotate entities from the chart data tables. This encompasses bar names and heights, axis labels and the chart title, as shown in (b) of Table 1. For each chart, we rank the bars by height, so *first** refers to the bar of rank 1, i.e. the highest bar. The bar name and height of the highest bar are labeled as *firstX* and *firstY* , respectively. Similarly, we label the second bar’s name (*secondX*) and its height (*secondY*), and so on until the lowest bar (*lastX*, *lastY*). The entities of the axes are annotated with separate labels as well: *X* for x-axis, *Y* for y-axis. The label *title* is used when the summary describes the topic of the chart by copying or paraphrasing its title.

Further we annotated text describing information that is inferred from the chart as **analytical labels**. Specifically, analytical information includes relations between datapoints or approximated values of bar heights. Following the observations in our dataset we suggest these 5 analytical categories: addition, multiplication, grouping, slope, height approximation. In the addition relation (*addition_**), two or more bars are compared given their height, stating the difference in heights. Similarly, the multiplication relation (*multiplication_**) states the difference coefficient.

Instead of describing the bars individually or in terms of their difference, bars are often divided into groups (*group_**) and referred to as such with the names or heights. The group height can be an average or a sum of heights included in the group.

Label names for addition, multiplication and grouping signal the bars they describe, for example *addition-1-4* labels the difference between the highest and fourth bar; *multiplication-2-3* labels the coefficient between the heights of the second and third bar, while *groupSecondLastX* and *groupSecondLastY* describe the names and heights of the grouped second and lowest bars.

When the entity on the x-axis is ordinal, summaries often describe the general slope with a numerical value. We label this as *slope*.

Bar heights are often not provided with their actual values in the summaries. We label their approximations with *approx**, for example *approx-firstY* refers to the approximated height of the highest bar.

3.4. Semantic Alignment between Charts and Summaries

We annotated the raw text of summaries by aligning the summaries to the elements from the chart using the

[1] https://www.prolific.co/

above mentioned basic and analytical labels. The first step of the aligning process was automatic relying on string matching and simple paraphrasing, both comparing the entity tokens in the chart data with the tokens in the summary. In the second step, the labeled summaries were revised manually.

The examples below show annotated full summaries (1a and 1b) and single sentences (2a and 3a) with color-coded basic and analytical labels.

1a This chart looked at causes of Obesity in Kiribati$_{title}$. 25%$_{lastY}$ was attributed to fast food$_{lastX}$ and 35%$_{secondY}$ to lack of exercise$_{secondX}$. The highest cause$_{X}$ was genetic$_{highestX}$ at 40%$_{highestY}$ which was nearly twice$_{multiplication-1-3}$ as attributable to fast food$_{lastX}$.

1b Causes of Obesity in Kiribati$_{title}$ are presented here. The biggest cause$_{X}$ is genetic$_{highestX}$ at 40%$_{highestY}$. However lifestyle factors$_{groupSecondLastX}$ are a larger contributor with fast food$_{lastX}$ and lack of exercise$_{secondX}$ combined shown to cause obesity in 60%$_{groupSecondLastY}$ of cases. This is broken down into lack of exercise$_{secondX}$ at 35%$_{secondY}$ and fast food$_{lastX}$ at 25%$_{lastY}$.

2a This is followed by Montevideo$_{secondX}$ at just under 30 °C$_{approx-secondY}$.

3a The median salary of women$_{Y}$ grew at a cost rate of approximately $5,000$_{slope}$.

3.5. Corpus Statistics

Table 3 shows the basic statistics. The corpus contains 47 bar charts that are based on 18 topics, for example *camera price per brand*, *average temperature per city*, *gender pay gap per country*, *stock price per day*. Nine topics have two charts, three topics have three charts, two topics have four, and further two topics have five charts, two topics have just a single chart. Each chart comes with 22 different summaries on average (min. 20, max. 23). Each summary has about 3 sentences and 54 tokens. In total, there are 581 summaries with a *neutral* message, 184 summaries with a *proportional* message, 183 with an *inverse*, and 115 with an *emphasis* one.

#topic	#chart	#summary	#token	#sentence
18	47	1,063	57,420	3,356

Table 3: Overall corpus statistics.

4. Properties of human-written chart summaries

We further explore analytical labels as well as the effect of different chart messages on the information ordering in the summaries.

4.1. Analytical information

We consider the occurrence of analytical categories: bar height approximations, bar grouping, multiplication, addition and slope. 67% of summaries include at least one analytical category. 53% of summaries have at least one height approximation, i.e. at least one of the described bar heights was rounded and not exact. 25% of summaries include an analytical category other than a height approximation. Table 4 presents the frequency of categories of analytical labels. Height approximations are by far the most common type of analytical information. Of all bar height references in the corpus 1,294 are approximated. Most of these are rounded to the nearest major tick mark on the y-axis. The rest of 62% height references is exact.

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)

Table 4: Frequency of analytical entities.

The second most frequent analytical category are group descriptions, either just by name or also height. We observe that group references most often occur when two or more bars are of similar height or when the bars are semantically closer than other bars (example in 1b with “lifestyle factors”).

Multiplication is a bit more frequent than addition when it comes to comparing bars with arithmetic operations. The values used to describe the multiplicative relations are mostly smaller integers between 2 and 5. For the additive relation, this varies: the summaries either include small integers or larger multiples of 5 (10, 25, 30).

The slope of the data is provided only 29 times. We note here that only 14 of 47 charts have an ordinal x-axis entity. For such charts, the workers also often provided other relations, such as grouping or addition.

4.2. Entity ordering given the intended message

We found that the order of entities (bar names and heights) in summaries varies with the intended message as shown in Table 5. The dominant narrative in summaries of charts with a *neutral* and *proportional* message is to follow the bar heights in descending order, starting with the highest bar, followed by the second one and so on. This is the case with 65.13% of *neutral* and 64.86% of *proportional* summaries. Some start with the lowest bar or other (usually the first bar along the x-axis).

In case of charts with an *inverse* message, 51.65% of their summaries start with the lowest bar, but in 39.01% of summaries participants chose to describe the highest

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

Table 5: The percentage of summaries starting with the highest, lowest or another bar given the intended message of the chart.

bar at the beginning of the summary even though the title prompts them to start with the lowest one. The inverse message might be more difficult to process, leading some participants to default to the descending order. Summaries of charts with an *emphasis* message start with the bar placed in focus in 71.27% of cases, while the rest of summaries start with another entity.

4.3. Human evaluation for summary preference

We conducted a small user study to explore desired properties of chart summaries according to the readers in a situation where they are presented with a chart and its summary. We compared the preference between summaries with basic facts versus those with additional analytical information.

Methodology. In the experiment, participants were shown a chart and two summaries describing it. They were asked to select the summary they preferred and explain their choice in a text box. There was a third option (*neither*) in case neither of the two summaries was preferred over the other. 28 native speakers of English from the UK were recruited via Prolific. The hourly pay rate was 14.42 GBP.

Materials. The items are taken from the Barch dataset. We used 14 charts. For each chart two summaries were selected with a similar length and style, but one containing only basic information about bar heights, while the other included analytical information, such as bar comparisons or group references.

Results. The preferences were divided equally between basic and analytical summaries with only 7% of responses choosing *neither*. A breakdown by participants shows that a big majority was not consistent in their choices, so they sometimes chose basic and sometimes analytical as the better summary. From the written explanations it is clear that the participants either want the summary to provide additional analytical insights as opposed to exactly repeating the same facts as the chart. Other responses state the participants prefer to think and analyze the chart data by themselves, so the summary should be basic and brief. Other responses comment on the preferred conciseness or descending order of information by heights. This study hence shows that the readers have diverse needs which might not be constant for the same person across different charts.

5. Generation of Summaries

One use case of our dataset is to train a data-driven generation system for automatically generating summaries of charts. We use three different neural sequence-to-sequence models on our dataset and analyze their performance. The input to the NLG systems is chart data, preprocessed according to the requirements of each model. The output is the chart summary.

5.1. Experimental Setup

We divide the Barch dataset into splits in two ways: 1) seen topics in the test set 2) two unseen topics in the test set. Each time we create a train, validation and test set. In the *seen* split, all topics appear in the train, validation and test set; however, the test set contains unseen charts of these seen topics. Note that there are multiple charts per topic (See 3.5). In contrast, the train and validation set in the *unseen* split contain the same topics, but the test set includes two topics that were not seen at training or validation. By using two different splitting methods we test model robustness against domain change.

Further, we experiment with the scope of information in the input: it either includes only basic data about the title, axis labels and the name and height of each data point, or it additionally includes analytical data about relations between bars. While the basic data is taken from the plotting step, we generate the analytical data based on basic data. For example, for the chart in Table 1, the value for the addition relation between the highest and second highest bar under the label *addition-1-2* is 5 ($40 - 35 = 5$).

5.2. NLG Architectures

We experiment with three neural sequence-to-sequence architectures to generate summaries from chart data. The implementation details and hyperparameters for each model are in the Appendix.

LSTM. We train a baseline LSTM encoder-decoder model with attention (Bahdanau et al., 2015). The input sequence is a linearized set of key-value pairs, where the keys are labels and the values are the chart data values. Training a Transformer (Vaswani et al., 2017) from scratch led to poor results.

C2T. Chart2Text (henceforth, C2T), a model by Obeid and Hoque (2020), is a Transformer architecture adapted specifically for generating summaries of charts from data tables. It does not use any pre-training. The authors train and test their model on their dataset.

We noticed that training on their data as well as on Barch worsens the performance in terms of input fidelity, so we use their model architecture to train on Barch only.

KGPT. The third model we use is KGPT (Knowledge-Grounded Pre-Training for Data-to-Text Generation) by Chen et al. (2020). In contrast

Test domain	Input data	Model	BLEU	BERTSCORE-F1	GPT-2 ppl.	grammaticality	support (%)
seen	basic	LSTM	6.42	0.84	65.36	3.75	25.00
		C2T	20.85	0.87	66.06	4.30	80.00
		KGPT	**32.96**	**0.89**	18.71	**6.55**	37.21
	basic +analytical	LSTM	8.19	0.84	66.92	3.45	24.49
		C2T	26.16	0.88	41.05	4.90	**94.74**
		KGPT	19.90	0.86	**7.27**	5.10	38.89
unseen	basic	LSTM	19.84	0.86	62.44	3.17	7.14
		C2T	24.04	0.86	75.55	2.25	50.00
		KGPT	21.11	0.87	16.32	6.30	19.05
	basic +analytical	LSTM	20.57	0.86	123.98	3.08	11.76
		C2T	30.03	**0.89**	28.48	2.70	80.77
		KGPT	22.93	0.87	12.40	4.17	14.29

Table 6: Results from generation experiments with the LSTM baseline, C2T (Obeid and Hoque, 2020) and KGPT (Chen et al., 2020). On the seen test domain, KGPT generates the most fluent and grammatical text, but the summaries of C2T are actually factual given the input chart. Using an unseen test domain results in a performance drop for all models.

to C2T, this is a Transformer-based model built for text generation from data in general. It has been pretrained on multiple datasets and its checkpoint is made available for fine-tuning. We fine-tune the sequence encoder on our data.

5.3. Evaluation of NLG

We evaluated the NLG outputs for the following task-relevant aspects: fluency and grammaticality, factuality (support) given the input chart and similarity to the references. We employed automatic as well as human evaluation.

We use multi-reference BLEU-4 (Papineni et al., 2002) and BERTSCORE (Zhang et al., 2020) for **surface similarity to human-written summaries**. We provide F1 scores for BERTSCORE. We use **perplexity**[2] from the GPT-2 model (Radford et al., 2019) as a proxy for text fluency.

In a rating experiment, human annotators assessed the **gramaticality** of generated summaries on a 1-7 Likert scale, with 1 marking a completely ungrammatical text and 7 a grammatically correct text. Each output summary was rated by two annotators.

For **support**, we manually split each generated summary into claims. We provided the annotators with the summaries and corresponding charts and asked them to judge the claims as true, false or not applicable given the chart. For example, a true claim about chart in Table 1 is *The chart shows obesity causes in Kiribati*. A false claim could be *The lack of exercise is the largest cause of obesity*. A claim that does not apply to the chart cannot be assessed for factuality, e.g. *Stress is a minor cause of obesity*, since the chart does not include information about stress. Again, each summary was evaluated for factuality by two annotators. We calculate the support score as the percentage of true claims (with annotators in agreement) given all claims.

6. Results

We evaluate the three NLG models (LSTM, C2T, KGPT) given the generated test summaries. We first discuss the quantitative metrics[3], followed by a manual inspection of the outputs.

6.1. Quantitative

Table 6 presents the evaluation metrics for the three models in the settings with the seen or unseen test domain as well as two different input data scopes. We find that BERTSCORE-F1 gives very similar values across all conditions and is hence not very informative.

In general, KGPT achieves the best scores on grammaticality and perplexity, but obtains rather poor scores for support, i.e. it generates fluent text that does not fit with the input data.

On the other hand, C2T has the best performance for support, but does worse on grammaticality and is also worse on fluency than KGPT. Overall, the LSTM baseline performs worst, with low scores on both support and grammaticality.

Comparing the test domains, reserving the test set to unseen topics gives higher BLEU scores, but it should be noted that the train set in this condition is also bigger while the test set is smaller. However, it is evident that the task is easier in the seen condition, as we can see from the generally higher grammaticality and support scores. In this condition, C2T outperforms KGPT in BLEU and still performs well in support, but at the cost of putting phrases together in an ungrammatical way.

We also observe that adding analytical data to the input improves BLEU for most models. In such settings, the models seem to learn how to map between the source and target better. This could be due to repetitions of words (for example, single bar names and group references) or because the model has access to additional

[2] https://huggingface.co/docs/transformers/perplexity

[3] The annotators agreed in roughly 93% cases on the factuality (true vs. rest)

information that was unaligned in the basic condition.

6.2. Qualitative

Table 7 provides generations of the three models for a seen topic with basic and analytical information in the input. A shown in Figure 2, the input chart title is *The most successful teams in a quiz competition*. It shows 6 teams (x-axis) given the number of mistakes they made (y-axis). The message of this chart is inverse.

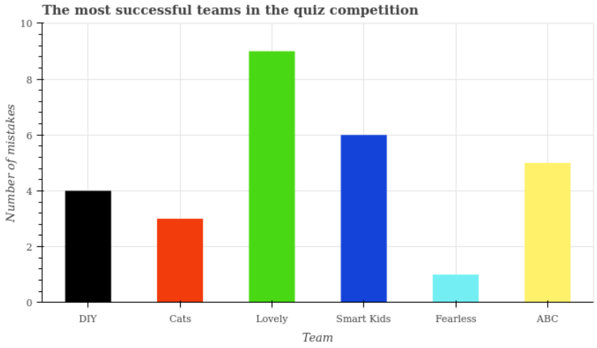


Figure 2: A bar chart with an inverse message.

LSTM
This chart shows the most successful teams in the quiz competition. The graph shows that the number of mistakes made made by most mistakes, followed by Fearless with 9 mistakes. Fearless was the least successful glacier with the next team team, with 1 mistake.
C2T
The chart shows the number of mistakes made per team in a quiz so that the highest bar shows the least successful - team Lovely with 9 number of mistakes, Smart Kids with 6, ABC with 5, DIY with 4, Cats with 3 and the only 1 number of mistakes made by Fearless.
KGPT
This chart shows the most successful teams in the quiz competition. The chart shows the most successful teams in the competition. The most successful team is the team representing the second highest group in the competition.

Table 7: Example outputs for seen topics with input data including basic and analytical chart information. We color-code non-factuality, copied train targets and off-topic vocabulary.

Both the LSTM and KGPT struggle with repetition of words or even entire sentences. Their summaries also contain off-topic words that come from other charts. The LSTM also generated a false claim. The summary by C2T does not struggle with these issues, however, it copies the entire summary from the target train set. The train set includes a chart with the same underlying data, but with a different title conveying a proportional message (*The least successful teams in the quiz competition*), from which the generated summary was copied.

6.3. Discussion

KGPT gives fluent and grammatical outputs because it was pretrained on extensive amounts of data-to-text corpora. However, it fails to generate input-related summaries. This indicates that fine-tuning on Barch was not successful, most probably because there are too few instances for that. Another possible reason for the poor performance of KGPT is the data discrepancy: in comparison to most data KGPT was trained on, Barch summaries span over multiple sentences, contain anaphoric references and discourse connectives, which are difficult to learn from only a handful of instances.

The C2T model has the opposite problem. Because it was trained only on our small dataset, it overfits to the train set and learns to copy parts of the train targets, often resulting in factual, but less grammatical outputs. Even when faced with unseen topics, it tries to apply the templates learned at training. Again, this is beneficial for support, but not for grammaticality.

The analysis of a sample of generations showed that the models did not consider the intended message when ordering information neither did they generate many analytical statements. These patterns were not learned from the data, however, there are ways to counter that. The intended message could be prepended to the input sequence as a special token or encoded in a trainable weight as a part of the encoder to condition the decoder on the intended message. A possible chart-to-text NLG model could also incorporate a content selection and planning component that constraints the decoder to generate entities in a certain order (Puduppully et al., 2019; Moryossef et al., 2019), which could increase the output diversity as well.

7. Conclusions

We present Barch, a dataset with human-written summaries of bar charts and annotated with basic chart information as well as analytical information inferred from the chart, e.g. approximated heights or relations between data points, as produced by human writers. This fills in a gap in the collection of chart-summary datasets, which have been either generated with templates (but including synthetically produced analytical statements in the summary) or humans (but not considering analytical statements).

Our analysis of Barch summaries has shown that the intended message in the chart title affects the entity order in human-written summaries. The default descending order by bar heights can be replaced with a different ordering if humans are prompted to do so by the message in the title.

The collected summaries also demonstrate that humans naturally draw analytical observations about chart data and produce them in text, mostly as approximations of single bar heights, but also as descriptions of bar groups or relations between them.

In a separate experiment on summary comprehension, we found evidence that readers sometimes prefer a chart summary with analytical facts over a summary that is a textual version of the chart data, containing only basic information. However, other properties of the text play a role in the choice. This calls for a more focused study that isolates the content preferences from style and length preferences.

A potential use case of Barch is training NLG systems. Our experiments with three NLG models suggest that, in contrast to human-written summaries, the generated summaries do not include analytical information and do not present the entities in various orders. On top of that, these models struggle either with fluency and grammaticality or input fidelity when trained on a small dataset like Barch.

One potential step towards improving chart-to-text generation is to enrich the datasets with analytical information and design the NLG model to generate such statements. Pooling the various chart-summary datasets and further exploring pre-training could counter the copying behavior as well.

8. Acknowledgements

We thank Rudy Khalil and Emeka Udeh for their contribution to the data collection process.

Appendix

Implementation details for NLG experiments. In the *seen* conditions, the sizes were 660/213/190 instances for the train/validation/test set, respectively. In the *unseen* conditions, the sizes were 787/139/137 for the train/validation/test splits, respectively.

If not stated otherwise, the implementation uses the default hyperparameter setting for C2T and KGPT.

The hyperparameters for the **LSTM** baseline implemented with OpenNMT (Klein et al., 2017): bidirectional LSTM encoder, unidirectional LSTM decoder, encoder embedding size 512, decoder embedding size 768, encoder and decoder RNN size 1024, 2-layer RNN in encoder and decoder, general global attention, using copy attention, batch size 64, dropout 0.3, Adam optimizer (Kingma and Ba, 2015), learning rate 0.001, 300 epochs. At inference: beam size 3, bigram block decoding.

The hyperparameters for training the **C2T** model: batch size 6, dropout 0.1, learning rate 0.0001, 80 epochs. At inference: beam size 6.

The hyperparameters for fine-tuning the **KGPT** sequence encoder: batch size 16, learning rate 2e-5, 30 epochs. At inference: beam size 2.

9. Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. January. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

Chen, W., Su, Y., Yan, X., and Wang, W. Y. (2020). KGPT: Knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8635–8648, Online, November. Association for Computational Linguistics.

Clive, J., Cao, K., and Rei, M. (2021). Control prefixes for text generation. *ArXiv*, abs/2110.08329.

Demir, S., Carberry, S., and McCoy, K. F. (2012). Summarizing information graphics textually. *Computational Linguistics*, 38(3):527–574.

Ferres, L., Verkhogliad, P., Lindgaard, G., Boucher, L., Chretien, A., and Lachance, M. (2007). Improving accessibility to statistical graphs: The IGraph-Lite system. In *Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '07, page 67–74, New York, NY, USA. Association for Computing Machinery.

Greenbacker, C., Carberry, S., and McCoy, K. (2011). A corpus of human-written summaries of line graphs. In *Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop*, pages 23–27, Edinburgh, Scotland, July. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*. ICLR.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Li, L., Ma, C., Yue, Y., and Hu, D. (2021). Improving encoder by auxiliary supervision tasks for table-to-text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5979–5989, Online, August. Association for Computational Linguistics.

Moryossef, A., Goldberg, Y., and Dagan, I. (2019). Step-by-step: Separating planning from realization in neural data-to-text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2267–2277. Association for Computational Linguistics.

Obeid, J. and Hoque, E. (2020). Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 138–147, Dublin, Ireland, December. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J.

(2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. ACL.

Puduppully, R., Dong, L., and Lapata, M. (2019). Data-to-text generation with content selection and planning. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, volume 33, page 6908–6915.

Pusse, F., Sayeed, A., and Demberg, V. (2016). Lingo-Turk: managing crowdsourced tasks for psycholinguistics. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 57–61, San Diego, California, June. Association for Computational Linguistics.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Talbot, J., Setlur, V., and Anand, A. (2014). Four experiments on the perception of bar charts. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2152–2160.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhu, J., Ran, J., Lee, R. K.-W., Li, Z., and Choo, K. (2021). AutoChart: A dataset for chart-to-text generation task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1636–1644, Held Online, September. INCOMA Ltd.

10. Language Resource References

Obeid, J. and Hoque, E. (2020). Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 138–147, Dublin, Ireland, December. Association for Computational Linguistics.

Zhu, J., Ran, J., Lee, R. K.-W., Li, Z., and Choo, K. (2021). AutoChart: A dataset for chart-to-text generation task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1636–1644, Held Online, September. INCOMA Ltd.