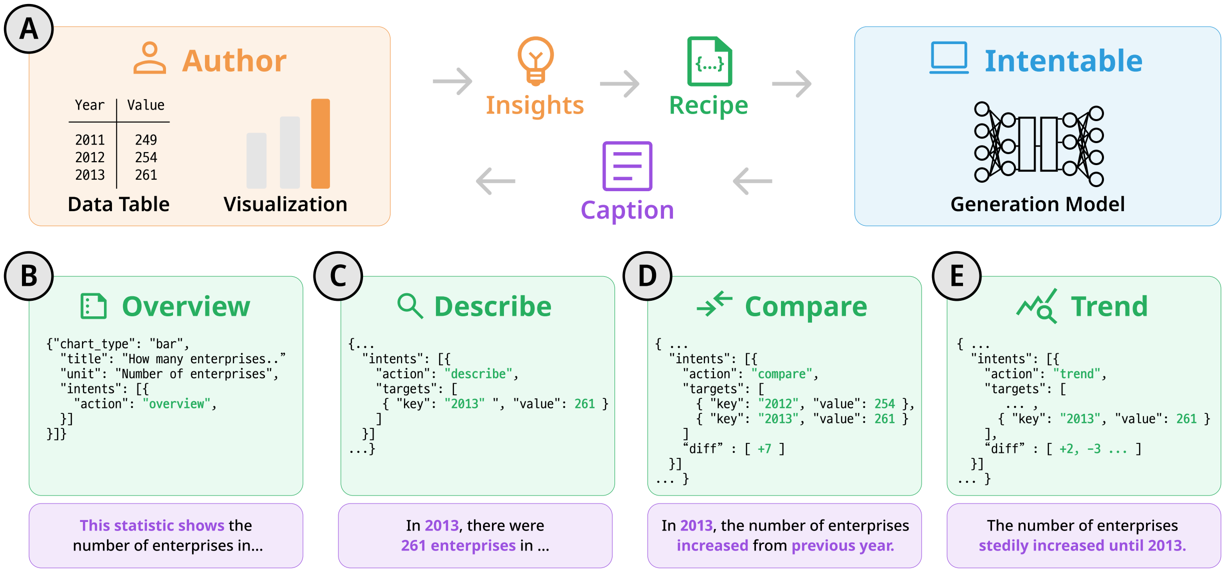
Intentable: A Mixed-Initiative System for Intent-Based Chart Captioning

Jiwon Choi[*]
Sungkyunkwan University

Jaemin Jo[†]
Sungkyunkwan University



```
{"chart_type": "bar",
  "title": "How many enterprises..”
  "unit": "Number of enterprises",
  "intents": [{
    "action": "overview",
  }]
}]}
```

This statistic shows the number of enterprises in...

```
{...
  "intents": [{
    "action": "describe",
    "targets": [
     { "key": "2013" ", "value": 261 }
    ]
  }]
...}
```

In 2013, there were 261 enterprises in ...

```
{ ...
  "intents": [{
    "action": "compare",
    "targets": [
      { "key": "2012", "value": 254 },
      { "key": "2013", "value": 261 }
    ]
    “diff” : [ +7 ]
  }]
... }
```

In 2013, the number of enterprises increased from previous year.

```
{ ...
  "intents": [{
    "action": "trend",
    "targets": [
       ... ,
      { "key": "2013", "value": 261 }
    ],
    “diff” : [ +2, -3 ... ]
  }]
... }
```

The number of enterprises stedily increased until 2013.

Figure 1: (A) Mixed-initiative interaction for caption authoring in Intentable. First, the author expresses their insights that they want to describe as caption intents. Together with contextual information (e.g., the title of a data table and a visualization type), the caption intents are encoded as a JSON-based caption recipe (code in the green boxes of B-E). Given a recipe, a generation model composes natural language sentences (sentences in the purple boxes of B-E). Intentable supports four intent types with different semantic levels: `Overview` (B), `Describe` (C), `Compare` (D), and `Trend` (E).

Abstract

We present Intentable, a mixed-initiative caption authoring system that allows the author to steer an automatic caption generation process to reflect their intent, e.g., the finding that the author gained from visualization and thus wants to write a caption for. We first derive a grammar for specifying the intent, i.e., a caption recipe, and build a neural network that generates caption sentences given a recipe. Our quantitative evaluation revealed that our intent-based generation system not only allows the author to engage in the generation process but also produces more fluent captions than the previous end-to-end approaches without user intervention. Finally, we demonstrate the versatility of our system, such as context adaptation, unit conversion, and sentence reordering.

Index Terms: Human-centered computing—Visualization—Visualization systems and tools; Human-centered computing—Human-computer interaction (HCI)—Interactive systems and tools

1 Introduction

We present a mixed-initiative caption authoring system, Intentable, that allows the author to steer an automatic caption generation process to reflect their intents. Captions, accompanied by visualizations offer various benefits to readers, such as explaining a visual mapping [3, 17, 39], emphasizing the author's takeaways [14], and providing better accessibility, especially for visually impaired readers [22]. However, authoring captions remains a time-consuming and demanding process as the author should analyze the data and compose the content manually.

With the rapid advance in natural language processing (NLP) technologies, several automatic systems [4, 9, 11, 24, 28] have been proposed to accelerate the authoring process by employing deep learning models. These systems aim to generate captions without user intervention given a chart image and/or a data table. However, such fully automatic systems often do not reflect the author's intent on captions, such as their insights on visualizations that they want to compose a caption for, due to the lack of user interaction. For example, there can be a visual element (e.g., a bar in a bar chart) and the corresponding row in the data table that the author wants to mention in the caption, but those systems do not provide a way to express such intent. Furthermore, although captions in the real world are very diverse in terms of the semantic content they deliver and the order of presentation [22], those systems lack the means to control such diversity, leaving it to a deep learning model.

Intentable overcomes these limitations by taking a mixed-initiative approach [8]. Our design goal is to take advantage of automatic caption generation while allowing the author to control it. Fig. 1A illustrates our mixed-initiative interaction. Through an interactive interface, the author first expresses their intent on captions, such as the content that each sentence should deliver and the order between them. The intent can be specified as a caption intent, high-level action and target pair. For each caption intent, the system synthesizes a natural language sentence considering contextual information (e.g., the title of the data table). We quantitatively and qualitatively evaluate Intentable and show that our intent-based

[*]e-mail: jasonchoi3@g.skku.edu
[†]e-mail: jmjo@skku.edu

approach not only produces more accurate captions than fully automatic approaches, but also offers flexibility in the generation process.

2 Related Work

Previous caption authoring systems can be classified into 1) storytelling systems, 2) rule-based generation systems, and 3) ML-based generation systems.

Storytelling systems allow the author to coordinate a written narrative with data visualizations to facilitate interpretation, so-called *narrative visualization* [32]. These systems do not generate the narrative itself, which is assumed to be given by the author, but help the author link the textual annotations with visualization elements expressively and flexibly [5, 30, 31]. Therefore, in storytelling systems, it is crucial to identify the mapping between annotations and visual elements, which has been facilitated by rule-based algorithms [13,23], neural networks [16,18,36], and crowd-sourcing [15]. Although these systems speed up the matching process, they do not generate narratives or captions themselves.

Besides storytelling systems, there have also been approaches to generating captions based on heuristics or templates. For example, iGRAPH-Lite [6] employs slot filling of short messages (i.e., templates) to produce captions that can help visually impaired readers. Another example is Voder [35], which automatically generates data facts and provides template-based textual descriptions of the data facts. Similarly, Wordsmith [1] uses “dynamic templates” to generate explanations for visualizations. More recently, AutoCaption [20] uses a set of 13 caption templates, each corresponding to a low-level analytic task such as trend and compare. However, these systems offer limited scalability as an expert has to prepare hand-engineered templates. Furthermore, the generated captions often lack diversity since they are based on a relatively small number of templates.

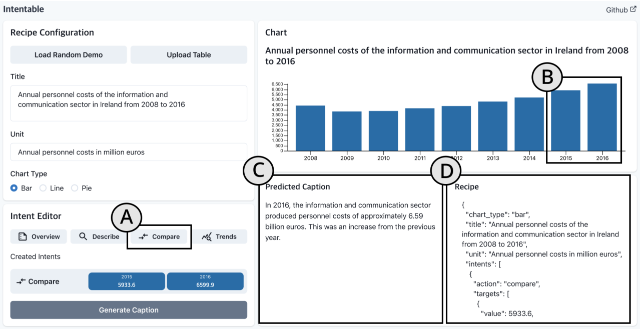
The development of deep neural networks enables the fully automatic generation of chart captions without human intervention. Inspired by the success of natural image captioning, various neural architectures have been adopted, such as CNN and LSTM encoder-decoder [4, 9, 10], encoder-decoder LSTM [34], and Transformers [24]. Most recently, Kantharaj et al. [11] collected 44,096 chart-caption pairs and compared the performance of several state-of-the-art architectures (e.g., T5 [29]). However, these end-to-end generation systems do not allow the author to engage in the generation process, neglecting the author’s intents on captions. In our work, we aim to support collaborative caption authoring where the author expresses their insights on visualization as caption intents, and the system composes the actual caption sentences considering the intents.

3 The Intentable System

In contrast to previous fully automated, thus monolithic, caption generation systems, Intentable allows the user to customize the output. Fig. 2 shows the user interface and a usage example of Intentable. Suppose an author, Jason, is inspecting the bar chart in Fig. 2 and wants to write a caption that compares the values of 2015 and 2016. He first chooses the type of his intent, i.e., *action*, Intentable supports four actions: `Overview`, `Describe`, `Compare`, and `Trend`. Since he wants to compare two values, he chooses `Compare` (Fig. 2A). Then, he selects the items that will be compared, i.e., *targets*, by clicking on the bars for 2015 and 2016 on the bar chart (Fig. 2B). Intentable encodes his intent as a caption recipe (Fig. 2D), and our generation model automatically composes caption sentences for the recipe (Fig. 2C). With our system, the user can customize 1) the number of caption sentences, 2) the order between the sentences, and 3) the content of each sentence.

3.1 Corpus Construction and Annotation

Our system generates captions considering a data table, a chart, and the author’s intent. Therefore, we first construct a training dataset where all three types of information are available. We crawled publicly accessible Web pages on Statista[1]. For each page, we collected the data table (including the title and unit), the visual encoding the chart used (e.g., chart type and the variables on the x and y axes), and the provided captions, which together, we call an **instance**. Our corpus consisted of 105,550 instances: 96,269 bar charts (including 16,790 multi-column bar charts, e.g., stacked bar charts), 7,620 line charts (including 1,749 multi-series line charts), and 1,661 pie charts. Note that the data table of every instance had at least one *key* variable (e.g., country name) and one *value* variable (the number associated with each country). This was about three times bigger than the corpus used in a recent study [11].



Title: Annual personnel costs of the information and communication sector in Ireland from 2008 to 2016

Unit: Annual personnel costs in million euros

Chart Type: ● Bar ○ Line ○ Pie

Created Intents: Compare — 2015 5933.6 — 2016 6599.9

Annual personnel costs of the information and communication sector in Ireland from 2008 to 2016

In 2016, the information and communication sector produced personnel costs of approximately 6.59 billion euros. This was an increase from the previous year.

```
{
  "chart_type": "bar",
  "title": "Annual personnel costs of the
information and communication sector in Ireland
from 2008 to 2016",
  "unit": "Annual personnel costs in million euros",
  "intents": [
   {
    "action": "compare",
    "targets": [
     {
      "value": 5933.6,
```

Figure 2: User Interface of Intentable

Since the captions were merely plain text without annotations, we applied rule-based heuristics to match a data value in the captions with a row in the data table. Each instance in our corpus had, on average, 2.52 sentences, with each sentence having 0.84 keys and 0.81 values. We attached the Python code we used as supplementary material.

3.2 Intent Tagging and Encoding

Since we want to generate intent-based captions, we need to identify 1) with what intent each caption in our corpus was made and 2) to which row in the data table (or visual element in the chart) the caption is related. To this end, we conducted a rule-based automatic tagging process inspired by the four-level framework by Lundgard and Satyanarayan [22], where semantic contents of captions were classified into four levels: 1) elemental and encoded, 2) statistical and relational, 3) perceptual and cognitive, and 4) contextual and domain-specific. The goal of the tagging process was to identify for each caption sentence the intent (among the four levels) and targets (i.e., the associated rows in the table).

We found that the first sentence of the captions usually does not contain values but reiterates the chart content (e.g., “This statistic shows the number of enterprises...”; Fig. 1B). We labeled these sentences as `Overview`, which corresponds to the first level (elemental and encoded) in the four-level framework.

About 39% of sentences simply described data facts of the data table, reading off one or multiple values in the table (“The sales were 1.2 trillion in 2020”; Fig. 1C). We classified these sentences as `Describe` and set the rows being described as targets. Among `Describe` sentences with two targets, there were sentences that made a comparison (“The sales increased from 1.2 trillion in 2020 to 1.5 trillion in 2021”). To distinguish such sentences from simple descriptions, we built a dictionary consisting of words frequently used for comparison, e.g., “increase” or “decrease.” We then separately labeled those sentences as `Compare` if at least one comparison word is included in the sentence (Fig. 1D). We could identify sentences with more than two targets and a comparison word but excluded those sentences from the training data because the intent of the caption was often multifold. Note that the sentences with `Describe`

[1] `https://www.statista.com/`

```
interface Recipe { chart_type: 'bar' | 'grouped_bar' |
'stacked_bar' | 'line' | 'multi_line' | 'pie';
title: string; unit: string; intents: Intent[]; }

interface Intent {
action: 'overview' | 'describe' | 'compare' | 'trend';
targets?: Target[]; /* empty when action is overview */
diff?: number[]; /* empty when action is overview or describe */}

type Feat = 'min' | 'max' | 'recent' | 'past';

interface Target {
features?: Feat[];
key: string; value: number;
series?: string; /* for multi-column data */ }
```

Figure 3: Grammar of Caption Recipes

or Compare actions could be seen as examples of the second level (statistical and relational) in the four-level framework.

We also found caption sentences that describe the overall trend in a chart ("Since 2012, the net capital stock has steadily increased."; Fig. 1E). Such Trend sentences were also identified using a dictionary, but we found most words in the dictionary conflict with the dictionary for the Compare action. To distinguish Trend from Compare, we checked 1) whether the data table has a temporal variable (e.g., year) and 2) whether at least one word about time (e.g., "until", "during", or "since") is present in the sentence. We labeled a sentence as Trend if both criteria were met. Trend sentences correspond to the third level (perceptual and cognitive) in the four-level framework.

Finally, we could identify 19% of sentences that do not belong to any of the labels we described. Most of these sentences provided background on data or subjective interpretation, which corresponds to the fourth level (contextual and domain-specific). We discarded these sentences as they were too subjective or often exploited background knowledge unavailable in the data table or chart.

After the tagging process, we obtained 190,830 intent-sentence pairs, including 72,564 Overview, 97,227 Describe, 5,961 Compare, and 15,078 Trend sentences. The dictionaries and code we used for tagging are available as supplementary materials.

Based on the result of the tagging process, we build one **caption recipe** for each instance in our corpus. Fig. 3 shows the JSON grammar we used to encode intents and the contextual information as a recipe. A recipe has the type of the chart, the title of the data table (e.g., "Percentage of COVID-19 vaccine doses administered worldwide by country income group"), the unit used in the table (e.g., "Percentage"), and a list of caption intents that we extracted from each caption sentence in the instance. Note that the actual caption sentences are not included in a recipe but are inferred from the recipe by the generation model. In contrast to a previous study [11], we choose not to provide the entire table as it could result in hallucination problems [11], and all necessary information to generate a caption (e.g., data values) is included as Target in the recipe.

We embed two types of extra information in the recipe. First, for Compare and Trend actions, we add a diff field that encodes the difference between two adjacent values in targets because we found that the model often works poorly on arithmetic operations (e.g., subtracting one value from another). The second information we embed is a features field. Since the model cannot access the full data table, it is unaware of whether the current target is at extremes (e.g., whether the target is the most recent data record or has the maximum value), although captions on extremes are frequent. Therefore, we explicitly provide such information for each Target. Example recipes can be found in the supplementary materials.

Table 1: Results of the Quantitative Evaluation

Model-Task-Size	BLEU	METEOR	BLEURT
BART-Table-base	**61.72**	63.85	0.177
BART-Table-large	59.95	**64.00**	0.156
BART-Intent-base	**62.53**	**66.09**	**0.285**
BART-Intent-large	57.77	62.40	**0.218**
T5-Table-small	59.45	63.18	0.187
T5-Table-base	60.30	63.73	0.196
T5-Table-large	60.00	63.58	0.146
T5-Intent-small	67.00	70.35	0.349
T5-Intent-base	**67.11**	**70.37**	**0.355**
T5-Intent-large	**67.31**	**70.41**	**0.351**

3.3 Model Architecture and Training

Inspired by the huge success of pre-trained language models in controllable text generation (CTG) [7, 12, 26], we chose to fine-tune two Transformer-based [37] encoder-decoder models, BART [19] and T5 [29]. Both models were trained to generate the original caption (i.e., the golden caption) given a caption recipe. In the training process, the corpus was split into 80%, 10%, and 10% of the data to create train, validation, and test sets, respectively. To obtain a model that is more robust in the order of caption sentences, we performed data augmentation by randomly removing a subset of intents and the corresponding sentences. As a result, we used 299,172 recipe-caption pairs for training. We chose the best checkpoint in terms of validation loss.

4 Evaluation

4.1 Quantitative Evaluation

The goal of the quantitative evaluation was to 1) understand the performance of different architectures (BART and T5) on intent-based caption generation tasks and 2) compare it to the previous end-to-end generation approach.

Models. We evaluated five variants of BART and T5 with different sizes to measure the capacity: **BART-base** (140M trainable parameters), **BART-large** (410M), **T5-small** (70M), **T5-base** (220M), and **T5-large** (770M).

Tasks. Each model was fine-tuned for two different tasks: **Table** and **Intent**. In the **Table** task, the model was given a recipe without the intents field but with a new data field that contains the list of the raw tokens in the data table. This task was similar to the generation task in a previous study [11] and corresponds to end-to-end caption generation without user intervention (i.e., no intents). On the other hand, in the **Intent** task, the model was given a caption recipe in Fig. 3, with intents but without the raw data table, which represents intent-based caption generation.

Metrics. We used BLEU [25, 27], METEOR [2], and BLEURT [33] as evaluation metrics. Traditionally, word-overlap-based metrics, such as BLEU and METEOR, have been widely used to measure the similarity between the generated text and the golden text. However, a recent study [21] suggested that these metrics are not highly correlated with human evaluation since they neglect the lexical context of a word. To overcome this issue, we also used BLEURT, which is known to be better correlated with human evaluation [33].

Results and Discussion. The benchmark result is shown in Table 1. Overall, all models achieved BLEU and METEOR scores near or higher than 60. The difference between architectures (BART and T5) and model sizes (small, base, and large) was not significant, and the performance was even degraded for larger models in the case of BART. This may indicate that even a small model had enough capacity for the generation tasks. However, we found a notable difference in tasks: all T5 models produced more accurate captions when intents were given (i.e., Intent tasks) than when data tables were given (i.e., Table tasks). The gain was about 7 points in the BLEU and METEOR scores and 0.17 points in the BLEURT score. This implies that intent-based generation not only supports user intervention but also produces more accurate captions. Furthermore, considering that the full data table is often unavailable in practice (e.g., only visualization is available), intent-based caption generation can be more ecologically valid as it only requires the data of target elements.

4.2 Qualitative Findings

In this section, we report on the qualitative findings of our model. All examples were generated using the T5-Intent-base model.

Context Adaptation. We conducted an ablation study on T5-Intent-base where we masked each field in Fig. 3 and inspected the loss in metrics. We found that ablating `title` and `unit` fields results in a loss of about 15 points in BLEU. This indicates that our model actually attends to those fields to determine the context of captions. For example, one can adapt captions for a different domain by changing the `title` and `unit` fields as follows:

> `title:` **Sales value** of **craft beer in Ontario, Canada** from 2012 to 2017
> `unit:` **Sales value** in million Canadian dollars.
>
> (Original) This statistic shows the **sales value** of **craft beer in Ontario, Canada** from 2012 to 2017. In 2017, ...

> `title:` **Sales value** of **computers in Seoul, Korea** from 2012 to 2017
> `unit:` **Sales value** in million Korean wons.
>
> (Adapted) This statistic shows the **sales value** of **computers in Seoul, Korea** from 2012 to 2017. In 2017, ...

We also found that when one of the two fields is missing, the missing information is inferred from the other field, as shown in the following examples. We highlighted the inferred parts in bold. Note that the inferred parts can be incorrect; for example, in the second example below, the model is assuming the data is about Apple's App Store, which is not provided in the data.

> `title:` **How expensive** is it to get a cat?, `unit:` (Ablated)
>
> This statistic shows the results of a survey on how expensive it is to get a cat ... 433 **U.S. dollars** was spent on buying a cat from a breeder.

> `title:` (Ablated), `unit:` **Share** of sales
>
> This statistic depicts **the share of digital and physical sales in Apple's App Store** ...

Unit Conversion. We found that our model learns the relationships between orders of magnitude (million-billion-trillion) and performs conversion between the units. In the example below, the target value was given in billions, but the model converted it to trillions and rounded off the value.

> `unit:` Amount in **billion euros**, `targets: [{value:` **2138.59**, ...}, ...]
>
> The largest value of insurance portfolio was found in 2015, when it reached **2.14 trillion euros.**

Relation Capturing. We provided the `diff` field that explicitly encodes the difference between adjacent target values to facilitate the comparison between them. See an example below:

> `targets: [{key:` **2017**, `value:` **33.5**}, `{key:` **2018**, `value:` **32.8**}],
> `diff:` **[-0.7]**
>
> In 2018, the sales value **decreased** to 32.8 million British pounds.

To see the impact of the `diff` field, we trained a model without this field and checked whether the model chooses a correct verb (i.e., "increase" vs. "decrease") for 1,600 randomly chosen `Compare` and `Trend` caption sentences. Without `diff`, the F1 score of classification was 0.76, while with `diff`, it increased to 0.88, which indicates that the model exploits the `diff` information in generating captions.

Sentence Reordering. Since our generation model is ordering-aware, the author can flexibly change the order between caption sentences by manipulating the order of the corresponding intents.

> `intents: [{action:` **overview**}, `{action:` **describe**, ...}]
>
> **This statistic shows** the results of a survey on the awareness of podcasts in Italy in 2019. According to data, **69 percent of respondents stated** that they knew what podcasts were.

> `intents: [{action:` **describe**, ...}, `{action:` **overview**}]
>
> During the survey period, **69 percent of respondents stated** that they knew what podcasts were. **This statistic shows** the results of a survey on the awareness of podcasts in Italy in 2019.

Through qualitative analysis, we could understand how the model attends to the information given in a recipe. We could also confirm the versatility of our system, such as context adaptation, unit conversion, and sentence reordering.

5 Limitations and Future Work

Semantic Accuracy. We found that our model sometimes attends too much to the contextual information, generating false captions. For example, when a temporal dataset is given, the model often attends too much to the most recent year even though it was asked to write a caption about another year. We believe that this issue can be alleviated by adopting another loss function that puts constraints on semantic accuracy [38] in the training process.

Generalizability. It would also be interesting to generalize the model for more than two targets or mixed intents. When we fed three targets with a `Describe` action, the model often ignored the third target or was confused about the number comparison. This could be addressed by augmenting the training data with more than two targets.

Diversity. Although Intentable produces diverse captions according to data contexts, the generation quality can be improved if more diverse training examples are given. Indeed, our training corpus mostly consisted of bar charts, which may limit the expressiveness of the model. We plan to build a larger corpus with richer features (e.g., more chart types or multiform visualizations) to improve the generation quality.

6 Conclusion

We present an intent-based caption authoring system, Intentable. We classify the sentences of the large-scale corpus we built into four intent types and derive a grammar to encode the intents as caption recipes. To generate caption sentences from recipes, two Transformer-based encoder-decoder architectures were fine-tuned and compared. In our quantitative and qualitative evaluation, we found that our intent-based caption generation produced more fluent captions than the previous fully automated approaches and demonstrated the versatility of our system.

Acknowledgments

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00421, AI Graduate School Support Program (Sungkyunkwan University)) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1F1A1073101).

REFERENCES

[1] Wordsmith by automated insights, inc. `https://automatedinsights.com/wordsmith`. Accessed: 2022-04-24.

[2] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

[3] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva. Beyond memorability: Visualization recognition and recall. *IEEE transactions on visualization and computer graphics*, 22(1):519–528, 2015.

[4] C. Chen, R. Zhang, E. Koh, S. Kim, S. Cohen, T. Yu, R. Rossi, and R. Bunescu. Figure captioning with reasoning and sequence-level training. 6 2019. doi: 10.48550/arxiv.1906.02850

[5] S. Chen, J. Li, G. Andrienko, N. Andrienko, Y. Wang, P. H. Nguyen, and C. Turkay. Supporting story synthesis: Bridging the gap between visual analytics and storytelling. *IEEE transactions on visualization and computer graphics*, 26(7):2499–2516, 2018.

[6] L. Ferres, G. Lindgaard, L. Sumegi, and B. Tsuji. Evaluating a tool for improving accessibility to charts and graphs. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 20(5):1–32, 2013.

[7] S. Goldfarb-Tarrant, T. Chakrabarty, R. Weischedel, and N. Peng. Content planning for neural story generation with aristotelian rescoring. *arXiv preprint arXiv:2009.09870*, 2020.

[8] E. Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 159–166, 1999.

[9] T.-Y. Hsu, C. L. Giles, and T.-H. K. Huang. Scicap: Generating captions for scientific figures. pp. 3258–3264, 10 2021. doi: 10.48550/arxiv.2110.11624

[10] S. E. Kahou, V. Michalski, A. Atkinson, Á. Kádár, A. Trischler, and Y. Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017.

[11] S. Kanthara, R. T. K. Leong, X. Lin, A. Masry, M. Thakkar, E. Hoque, and S. Joty. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*, 2022.

[12] M. Khalifa, H. Elsahar, and M. Dymetman. A distributional approach to controlled text generation. *arXiv preprint arXiv:2012.11635*, 2020.

[13] D. H. Kim, E. Hoque, J. Kim, and M. Agrawala. Facilitating document reading by linking text and tables. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pp. 423–434, 2018.

[14] D. H. Kim, V. Setlur, and M. Agrawala. Towards understanding how readers integrate charts and captions: A case study with line charts. *Conference on Human Factors in Computing Systems - Proceedings*, 1 2021. doi: 10.1145/3411764.3445443

[15] N. Kong, M. A. Hearst, and M. Agrawala. Extracting references between text and charts via crowdsourcing. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 31–40, 2014.

[16] C. Lai, Z. Lin, R. Jiang, Y. Han, C. Liu, and X. Yuan. Automatic annotation synchronizing with textual description for visualization. *Conference on Human Factors in Computing Systems - Proceedings*, 4 2020. doi: 10.1145/3313831.3376443

[17] S. Lallé, D. Toker, and C. Conati. Gaze-driven adaptive interventions for magazine-style narrative visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 27(6):2941–2952, 2019.

[18] S. Latif, Z. Zhou, Y. Kim, F. Beck, and N. W. Kim. Kori: Interactive synthesis of text and charts in data documents. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):184–194, 2021.

[19] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[20] C. Liu, L. Xie, Y. Han, D. Wei, and X. Yuan. Autocaption: An approach to generate natural language description from visualization automatically. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 191–195. IEEE, 2020.

[21] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.

[22] A. Lundgard and A. Satyanarayan. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE transactions on visualization and computer graphics*, 28(1):1073–1083, 2021.

[23] R. Metoyer, Q. Zhi, B. Janczuk, and W. Scheirer. Coupling story to visualization: Using textual analysis as a bridge between data and interpretation. In *23rd International Conference on Intelligent User Interfaces*, pp. 503–507, 2018.

[24] J. Obeid and E. Hoque. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. *INLG 2020 - 13th International Conference on Natural Language Generation, Proceedings*, pp. 138–147, 10 2020. doi: 10.48550/arxiv.2010.09142

[25] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[26] A. P. Parikh, X. Wang, S. Gehrmann, M. Faruqui, B. Dhingra, D. Yang, and D. Das. Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*, 2020.

[27] M. Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191. Association for Computational Linguistics, Brussels, Belgium, Oct. 2018. doi: 10.18653/v1/W18-6319

[28] X. Qian, E. Koh, F. Du, S. Kim, J. Chan, R. A. Rossi, S. Malik, and T. Y. Lee. Generating accurate caption units for figure captioning. In *Proceedings of the Web Conference 2021*, pp. 2792–2804, 2021.

[29] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[30] D. Ren, M. Brehmer, B. Lee, T. Höllerer, and E. K. Choe. Chartaccent: Annotation for data-driven storytelling. In *2017 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 230–239. Ieee, 2017.

[31] A. Satyanarayan and J. Heer. Authoring narrative visualizations with ellipsis. In *Computer Graphics Forum*, vol. 33, pp. 361–370. Wiley Online Library, 2014.

[32] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics*, 16(6):1139–1148, 2010.

[33] T. Sellam, D. Das, and A. P. Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020.

[34] A. Spreafico and G. Carenini. Neural data-driven captioning of time-series line charts. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pp. 1–5, 2020.

[35] A. Srinivasan, S. M. Drucker, A. Endert, and J. Stasko. Augmenting visualizations with interactive data facts to facilitate interpretation and communication. *IEEE transactions on visualization and computer graphics*, 25(1):672–681, 2018.

[36] N. Sultanum, Z. Bylinskii, and Z. Liu. Leveraging text-chart links to support authoring of data-driven articles with vizflow. *Conference on Human Factors in Computing Systems - Proceedings*, 5 2021. doi: 10.1145/3411764.3445354

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[38] Z. Wang, X. Wang, B. An, D. Yu, and C. Chen. Towards faithful neural table-to-text generation with content-matching constraints. *arXiv preprint arXiv:2005.00969*, 2020.

[39] Q. Zhi, A. Ottley, and R. Metoyer. Linking and layout: Exploring the integration of text and visualization in storytelling. In *Computer Graphics Forum*, vol. 38, pp. 675–685. Wiley Online Library, 2019.