Text-to-Viz: Automatic Generation of Infographics from Proportion-Related Natural Language Statements

Weiwei Cui, Xiaoyu Zhang, Yun Wang, He Huang, Bei Chen,
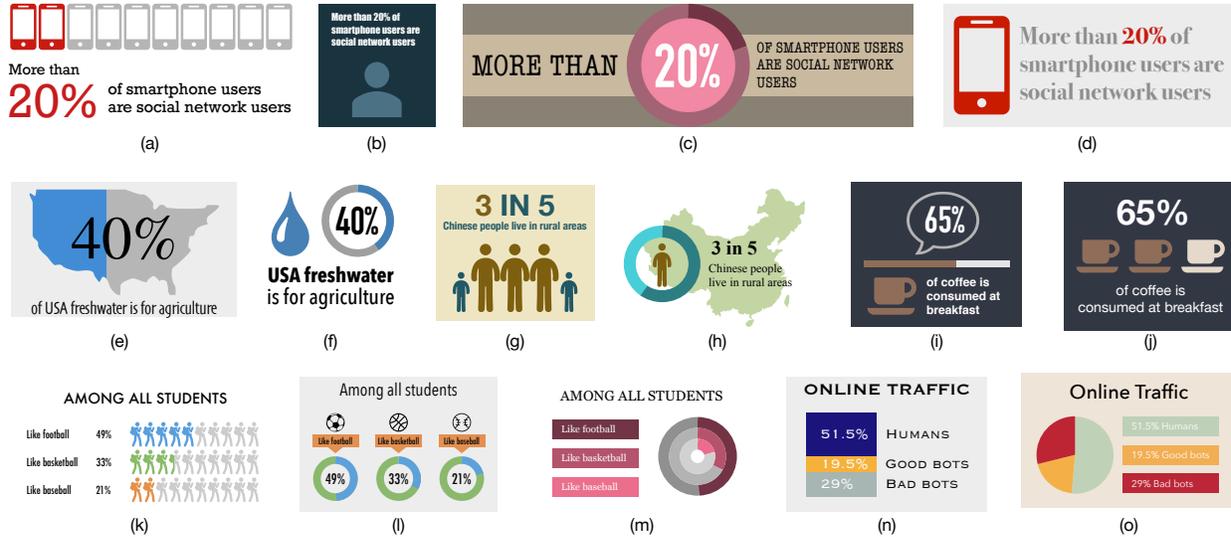Lei Fang, Haidong Zhang, Jian-Guan Lou, and Dongmei Zhang

Figure 1. Examples created by Text-to-Viz. (a)-(d) are generated from the statement: "More than 20% of smartphone users are social network users." (e) and (f) are generated from the statement: "40 percent of USA freshwater is for agriculture." (g) and (h) are generated from the statement: "3 in 5 Chinese people live in rural areas." (i) and (j) are generated from the statement: "65% of coffee is consumed at breakfast." (k)-(m) are generated from the statement: "Among all students, 49% like football, 32% like basketball, and 21% like baseball." (n) and (o) are generated from the statement: "Humans made 51.5% of online traffic, while good bots made 19.5% and bad bots made 29%."

Abstract— Combining data content with visual embellishments, infographics can effectively deliver messages in an engaging and memorable manner. Various authoring tools have been proposed to facilitate the creation of infographics. However, creating a professional infographic with these authoring tools is still not an easy task, requiring much time and design expertise. Therefore, these tools are generally not attractive to casual users, who are either unwilling to take time to learn the tools or lacking in proper design expertise to create a professional infographic. In this paper, we explore an alternative approach: to automatically generate infographics from natural language statements. We first conducted a preliminary study to explore the design space of infographics. Based on the preliminary study, we built a proof-of-concept system that automatically converts statements about simple proportion-related statistics to a set of infographics with pre-designed styles. Finally, we demonstrated the usability and usefulness of the system through sample results, exhibits, and expert reviews.

Index Terms—Visualization for the masses, infographic, automatic visualization, presentation, and dissemination.

---

1 INTRODUCTION

Information graphics (a.k.a. infographics) is a type of data visualization that combines artistic elements (e.g. clip arts and images) with data-driven content (e.g. bar graphs and pie charts) to deliver information in an engaging and memorable manner [24].Due to these advan-

- W. Cui, Y. Wang, H. Huang, B. Chen, L. Fang, H. Zhang, J. Lou, and D. Zhang are with Microsoft Research Asia. Emails: {weiweicu, wangyun, rayhuang, beichen, leifa, haizhang, jlou, and dongmeiz}@microsoft.com.
- X. Zhang is with the ViDi Research Group in University of California, Davis, and this work was done during an internship at Microsoft Research Asia. Email: xybzhang@ucdavis.edu.

tages, they are widely used in many areas, such as business, finance, and health-care, for advertisements and communications. However, creating a professional infographic is not an easy task. It is a time-consuming process and also often requires designer skills to ensure the perceptual effectiveness and aesthetics.

Much research has been devoted to investigating the design aspect of infographics [12, 27, 63] and developing authoring tools [30, 58, 69] to facilitate the creation of data-driven infographics. Based on different considerations, these authoring tools all strive to reach a balance between the ease-of-use and the power of features, and then to speed up the authoring process. However, these tools generally target advanced users. With complicated editing operations and technical concepts, these tools are not friendly to casual users, who, we believe, form another major category of infographic creators, other than professionals, such as graphic designers and data scientists [36].

Consider a hypothetical example in which a program manager, Nina, is preparing a presentation for her manager and wants to emphasize in her slides that "40% of US kids like video games." She

decides to add an infographic next to the statement with an authoring tool, e.g., DDG [30] or InfoNice [69]. Since Nina is not a professional graphic designer, she first needs to spend time (re-)familiarizing herself with the tool, such as the user interface and work flow. However, even if she were familiar with the tool, she still may not know where to begin to create a desired infographic, because all the existing authoring tools assume that the users have a clear or rough idea of what the final infographic may look like. Unfortunately, Nina has no design expertise and has little to no knowledge of how a professional infographic would look like. Therefore, she likely needs to go through existing well-designed infographics (in books or on the Internet) to look for inspiration. Based on the examined samples, she then settles on a design choice that has the best "return of investment" in terms of authoring time and her purpose of emphasizing the message.

From this example, we can summarize some common patterns for this user category. First, creators in this category only occasionally create infographics and thus are not proficient in the authoring tools. Second, they do not aim for the most creative/impressive infographics, which often involve complex designs and a long authoring time and are unnecessary in terms of "return of investment". Instead, something effective but professional is often sufficient for their purposes. Third, they often only have little design expertise, and would likely be unclear on how to design a decent infographic from scratch. On the other hand, if they were provided with good and relevant samples, they could often quickly pick one based on their personal preferences.

To address the needs of users in this category, we explored a new approach: to automatically generate infographics based on text statements. Since text is the most common form of communication to exchange information, we believe that this approach can help more people take advantage of infographics. To achieve this goal, there are two major challenges to overcome. The first one is to understand and extract appropriate information from a given statement. The second one is to construct professional infographics based on the extracted information. For the text understanding challenge, we first collected a corpus of real-world samples. Then, these samples were manually labeled to train a CRF-based model to identify and extract information for infographic constructions. For the infographic construction challenge, we analyzed and explored the design space of infographic exemplars that we collected from the Internet. Based on the design space, we proposed a framework to systematically synthesize infographics.

Considering the numerous types of information that can be represented by infographics [49, 62], and the numerous ways to express the same information textually and visually, it is impossible to cover the entire space in one paper. Instead, we decided to focus on a relatively small and isolated text-infographic space and build a proof-of-concept system for it. To achieve this goal, we first conducted a preliminary survey on the existing infographics to identify a category of information that is commonly represented by infographics and also has clear textual and visual patterns to process systematically. Based on the preliminary survey, we chose a subtype of information related to proportion facts (e.g., "Less than 1% of US men know how to tie a bow tie.") and built an end-to-end system to automatically convert simple statements containing this type of information to a set of infographics with pre-designed styles. Finally, we demonstrate the usability and usefulness of the system through sample results, two public exhibits, and expert reviews.

2 RELATED WORK

2.1 Information Graphics

Infographics are a type of data visualization dedicated to presentations and communications, rather than data exploration [33, 45]. By combining data content with visual embellishments, they can effectively deliver complex messages in a memorable and engaging manner [7, 23, 24].

Traditional research in infographics mainly focuses on the effects of visual embellishments and understanding their role as an attention-getting device [11, 24, 63].Although Tufte [67] argued that visual embellishments might be harmful, many other studies [10, 11, 12, 46] have shown that appropriate embellishments, such as pictographs [23],

annotations [53], and imageries [7, 16], do not seem to interfere with the correct perception of information, and can increase the long-term memorability of visualizations. For example, Haroz et al. [23] demonstrated that embellished charts are equivalent to plain charts in terms of reading speed and accuracy, but the added visual elements make them more memorable. Recently, studies have been conducted to directly understand infographics. For example, Bylinskii et al. [15] adopted OCR techniques to assign hashtags to infographics for information retrieval purposes. Madan et al. [39] used computer vision techniques to detect icons in infographics and proposed a solution to automatically summarize infographics.

Although these studies have extensively demonstrated the value of infographics from various perspectives, and started to investigate infographic understanding in general, none of them shares the same goal of ours, which is to build data-driven infographics automatically for general users.

2.2 Natural Language Interactions

Research on natural language processing [71] provides effective ways to analyze texts. Syntax analysis [20] and semantic analysis [8] can be used to discover hierarchical structures and understand meanings in text, but they are more appropriate for general language understanding problems. Sequence tagging tasks, such as Part-Of-Speech tagging [40] and Named Entity Recognition (NER) [48], aim to assign a categorical label to each word in text sequences, after that, text segments of different types can be obtained based on assigned categorical labels. Since the input of our system requires extracting text segments, we adopt the sequence tagging techniques. For sequence tagging techniques, rule-based methods [52] are straightforward, but costly and limited. There are also many machine learning methods proposed which are more robust and easier to generalize, including Hidden Markov Models (HMM) [9], Support Vector Machines (SVM) [4] and Conditional Random Fields (CRF) [34].

Recently, human languages have been used as an interface for visual data analysis. Systems like Articulate [66], DataTone [22], and Eviza [60] adopt various methods to transpile natural language queries to formal database queries, such as SQL, and extract analysis results accordingly. Although similar to these solutions in terms of understanding user inputs, our system does not relate to database querying. Instead, we directly visualize user inputs. Specifically, we extract essential semantics from user inputs and translate them into infographic representations for visual data storytelling. In our prototype, we have developed CRF with Convolutional Neural Networks (CNN) because of its high accuracy and efficiency of prediction.

2.3 Auto-Generated Data Visualizations

A group of visualization tools enables automatic generation of data visualization through a set of chart templates, including Tableau[1], Power BI[2], Many Eyes [68], and Voyager [72]. Analysts can specify subsets of data to generate visualizations directly. These tools have eased the exploration of data through quick chart generation. Although these tools allow users to easily generate visualizations within a few mouse clicks, users still have to make decisions about what charts to show and how data items are encoded visually.

More recently, machine learning models are leveraged to improve the quality of generated visualizations. These solutions may adopt machine learning techniques at different stages of the process. Some of them, such as DataSite [19], Foresight [21], and Voder [64], try to automatically extract significant insights from raw datasets, while other ones, such as Show Me [38], DeepEye [37], Draco [47], and VizML [26], use algorithms to identify the best visual form for a given data pattern. Although all these systems have reduced the effort required for creating data queries and visual encodings, the generated data visualizations are usually standard charts, which are not very expressive and customizable. In contrast to these systems, we have dif-

---

[1] https://public.tableau.com

[2] https://powerbi.microsoft.com/

ferent focuses at both stages, which is to convert the information that is explicitly provided by users to infographic-style visualizations.

To improve the expressiveness of generated standard charts, some systems also incorporate algorithms to automatically add visual cues [32]. For example, Contextifier [28] and Temporal Summary Images [14] support automatic generation of annotations. In a broader sense, most visual analytics systems also can automatically generate visualizations (often customized) from their targeting datasets for users to explore and analyze. However, these tools all require specific types of data format and aim at professional users trying to reveal complex patterns.

Our system also falls into this category. However, comparing with these existing tools that aim at standard charts or are deeply coupled with strict data formats, our system has very different input and output, helping users who know exactly what information to visualize and help them convert the information into infographics effortlessly.

2.4 Infographic Authoring Tools

Numerous interactive design tools have been proposed to help author more creative and customized infographics. For example, Lyra [58], iVisDesigner [54], and Charticulator [55] improve the flexibility of visualization creation through easy manipulation of low-level graphical specifications or declarative models with graphical user interfaces. iVoLVER [42] and Data Illustrator [36] design easy data binding and direct manipulation of graphical widgets to help users automatically transform visual elements to reflect underlying datasets.

To create more customized and engaging infographics, hand-drawn or uploaded images are introduced to improve visualization authoring systems. For example, Data-Driven-Guides [30] supports vector drawing and data binding, thus providing an authoring environment for data-driven infographics. DataInk [73] enables the creation of infographics through data binding with manipulation of pen and touch interactions. InfoNice [69] allows general users to convert unembellished charts into creative visualizations by manipulating visual marks and creating data-driven infographics easily.

Although these systems have greatly reduce the efforts of creating novel data-driven infographic designs they all require users to be familiar with certain non-trivial concepts, such as *vector*, *layer* and *data-binding*, which are often technique barriers for general users to master these systems. In addition, users need to have an initial idea of the final infographic design and realize the idea step by step, which is also challenging for inexperienced users. Therefore, these tools, by design, target at professional users, such as designers and data scientists.

In contrast to the existing authoring tools, our system specifically targets users without design experience. By dramatically reducing the complexity of the authoring process, with the cost of flexibility, our system can help general users create a variety of infographics automatically, saving them time and inspiring their creativity.

3 PRELIMINARY SURVEY OF INFOGRAPHICS

The goal of this survey was to better understand how infographics are used in real life and identify a specific type on which to build our proof-of-concept system. Following the methodology used by Segel and Heer [59], we first harvested a representative corpus of infographics from the Internet. We used "infographic" as a search keyword and downloaded the first 200 distinct results manually from Google Image Search. Due to Google's ranking mechanism, the top query results were believed to have good quality and relevance. Since in each search result, multiple related infographics may be arranged together for a comprehensive story, we further manually broke them down into 983 individual infographic units (Figure 2) based on the following criteria. A valid infographic unit should:

- deliver at least one message;
- contain at least one graphical element;
- be visually and semantically coherent and complete;
- not be able to split into smaller units that fulfill the above criteria.

We then categorized all the resulting infographic units into four main groups (Table 1) proposed by Artacho-Ram et al. [3]:
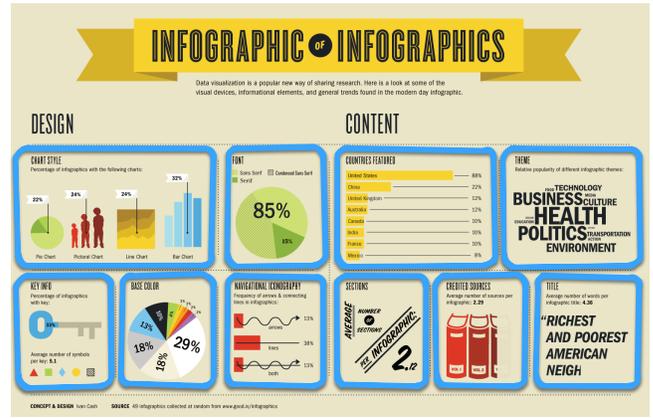

Figure 2. Example of breaking a search result (Infographic of Infographics [17]) down into individual valid infographic units. All valid units are marked with blue rectangles.

Category		Percentage
Statistics-based	Proportion	45.7%
	Quantity	28.3%
	Change	8.5%
	Rank	5.8%
	Other	6.2%
Timeline-based		1.5%
Process-based		1.1%
Location-based		0.7%
Other		2.2%

Table 1. Information categories of the collected infographics. Please note that the percentages are based on 983 infographic units, rather than 200 infographic sheets. Therefore, the numbers for non-statistics based infographics are exceptionally low, since they often take a larger space than statistics-based ones.

- **Statistical-based**: This is the main category of infographics, which commonly includes horizontal bar charts, vertical column charts, pictographs, and pie/donut charts, for summarizing a large amount of statistical information.
- **Timeline-based**: This category aims to show information and events happening over time, helping audiences realize chronological relationships quickly. Common visual representations include time-lines, tables, etc.
- **Process-based**: This category aims to instruct readers to take step-by-step actions to achieve a certain goal. They are often used to illustrate cooking recipes in magazines, or to clarify operations in workspaces or factories.
- **Location-based**: Infographics in this category generally contain a map and other annotations such as icons, diagrams, arrows, and tables for navigation. They are usually designed for places such as tourist spots, malls, factories, etc.

Please note that this categorization and the others discussed in this paper were first independently performed by three of the co-authors, and then documented as discussion results among the co-authors.

When examining infographics in the dominant statistics-based category, we further discovered diverse patterns in terms of graphic designs and the underlying messages, which were categorized into four major sub-categories, namely *proportion*, *quantity*, *change*, and *rank*:

- **Proportion**: This category of infographics aims to convey statistical information about how much a part occupies the whole, e.g., "More than *1/3* of the U.S. adult population is obese". According to our survey, this is the most common type of information de-
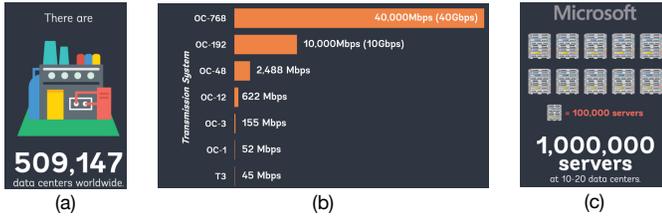
Figure 3. Exemplars of quantity-related infographics [44]: (a) embellishment icons, (b) horizontal bar charts, and (c) pictographs.

Figure 4. Exemplars of change-related infographics [57, 65]: (a) contrast color + side-by-side comparison and (b) contrast color + arrows.

Figure 5. Exemplars of rank-related infographics [56, 70]: (a) highlighted keyword + star embellishment and (b) ordered small-multiples.

livered in the data collected. Infographics in this category often have conspicuous visual and textual patterns. Visually, they often contain bar charts, donut charts, pie charts, pictographs, etc. Textually, the key information is often expressed in forms like "*n%*", "*m* in *n* ", "*m/n*", or "half of . . .". Because of its prevalence in practice and relatively prominent visual and textual features, we decided to build the proof-of-concept system for this infographic category.

- **Quantity**: The second largest category is related to quantities. These infographics emphasize the quantitative amount of a certain unit, such as income, population, or speed. Since the values cannot be described as fractions, visuals such as donut charts and pie charts do not apply here. Popular charts in this category include pictographs, vertical column charts, and horizontal bar charts (Figure 3(b) and (c)). However, we also noticed that more samples in this category do not incorporate any data-driven visuals. Instead, these infographics mainly contain embellishment visuals (e.g., icons) along with the quantity values highlighted with color, font size/weight, etc. (Figure 3(a)).
- **Change**: The third category aims at describing changes. Although values in this category are often also expressed as proportions or quantities, this category is different from the previous two by emphasizing the concept of change, such as "increase", "decrease", or "drop". In the corresponding infographics, arrows (Figure 4(b)), contrasting colors (Figure 4(a) and (b)), and side-by-side comparison (Figure 4(a)) are often employed in this category to depict the directions of changes.
- **Rank**: This category shows the relative position of a data item in a group, which is easy to identify based on ordinal numbers or specific symbols, such as "#", "No." or "Top *n*". However, the visualizations for single or multiple ranks are different. For a message that only contains one piece of ranking information, such as "Florida has the **3rd** largest homeless population," infographics usually highlight the key words, such as "*3rd*", and sometimes incorporate metaphoric embellishments, such as stars, medals, or trophy cups (Figure 5(a)). On the other hand, for a message that involves the ranking of multiple data items, infographics often arrange representing icons in an ordered list to directly convey their relative ranks (Figure 5(b)).

In addition to the information categories, we also discovered that the number of statistical facts covered in one infographic unit may vary. In many cases, an infographic unit only covers one fact. But there is also a considerable amount of infographics that contain multi-
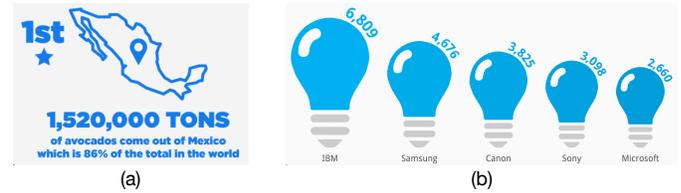
ple facts. Therefore, based on the number of facts and the relationships between facts in an infographic, we also identified four different categories, namely, *single*, *composition*, *comparison*, and *accumulation*, accounting for 32%, 25%, 27% and 16% of the collected infographic units respectively.

- **Single**: There is only one statistical fact in an infographic, it describes one facet of a subject. This is the simplest but the most commonly found type in our dataset. Since there is only one fact, the corresponding visualization is generally decided by its semantic type (Table 1).
- **Composition**: In this category, an infographic depicts more than one facet of a subject to form a complete picture of it. For example, in the statement, "In the United States alone, there were 10.5 billion searches in July 2009, which is a 106% increase from 5.1 billion in Dec. 2005," all of the numbers are used to provide information about searches in the United States. Since this category may involve different types of statistical facts, we do not cover it in our prototype system.
- **Comparison**: For this category, multiple facts are provided to compare the same facet of different subjects. Taking the statement "49% of students like football, while 33% of students like basketball" as an example, numbers "49%" and "33%" compare the students who like football or basketball. The facets of these two numbers are the same, i.e., students. Therefore, they are either combined into a bar chart or placed side-by-side using the same type of visual element. When visual elements are placed side-by-side, distinguishing colors, sizes, or icons are usually used to show difference and assist in comparison.
- **Accumulation**: Similar to *comparison*, this type of facts also depicts one single facet for different subjects. But these facts can be logically combined to form a larger whole. An example for *accumulation* is this statement, "60% of participants come from the US, while 40% come from Canada." In many cases, the numbers add up to one. And because of this characteristic, designers prefer to combine all the data into one pie chart, donut chart, or stacked bar chart, instead of visualizing individual facts separately.

Please note that the visualizations may overlap for categories *comparison* and *accumulation*. For some messages in the category of *accumulation*, designers may still use visualizations for *comparison*, e.g., side-by-side visuals, to emphasize their difference. Therefore, the visualizations suitable for *comparison* are also suitable for *accumulation*. However, the same is not true vice versa.

4 PROPORTION-RELATED INFORMATION

According to our preliminary survey, we decided to target proportion-related facts because of their dominance in our harvested dataset and distinct visual and textual patterns. In this section, we further analyze the textual and visual spaces of this type of information, on which we can build a model to automatically convert natural language statements into proper infographics.

4.1 Text Space

One goal of our system is to hide all technical concepts/details, so that the learning curve is minimized and everyone can easily create infographics without obstacles. Therefore, it is critical to allow users to provide the information to visualize using their daily language, rather

than formal declarative languages, such as JSON or XML. However, since a message may have many different ways to express in the natural language, we need to collect real-world samples to understand how people usually deliver proportion facts in text.

From a search engine index, we collected a dataset of real-world PowerPoint files that contains approximately 100,000 slides. From the slides, we used regular expressions to capture 5,562 statements that contain common expressions for fraction, such as *n%*, *m* in *n*, *m* out of *n*, and *half of*. Please note that not all the captured statements can be classified as *proportions* based on the above definitions, since statements such as "Company revenue increased by 24%" are also included here, which, obviously, should belong to *change*. Therefore, from the collected statements, we further manually sampled 800 valid proportion-related ones, and used them as the training dataset to build a machine learning model for text processing (Section 5.1).

At first, we also collected samples from online news articles. However, comparing statements from these two sources, we found that statements in PowerPoint slides are often shorter and more concise than those in news articles. Since PowerPoint-style statements are more focused on statistical facts and better match descriptions in infographics, we eventually removed samples from news articles and only kept the ones collected from PowerPoint slides.

4.2 Visual Space

To understand the infographics of proportion facts, we held a one-hour discussion with two experts, both of whom have at least four years of experience as freelance graphic designers. During the interview, we presented examples of proportion-related infographics, and tried to understand how these infographics were generated from scratch based on their experiences. The discussion resulted in a design space with four dimensions: *layout*, *description*, *graphic*, and *color*. Although there are a wide variety of designs, every infographic can be described as a tuple of these values. And the design space also aligns with the designers' mental model to create infographics. By thoroughly examining the collected samples, we summarized each dimension as follows.

4.2.1 Layout

Based on the existing layout taxonomies on general visualization design [5, 59], we analyzed and discussed the layout of infographics for proportional facts.

Layout for Single Fact   Infographics with single proportion facts are the simplest and most common type based on our preliminary survey. They are generally composed of two main visual elements: descriptions and graphics.

In most cases, the descriptions and graphics are arranged in a grid layout. They can be aligned in a horizontal (Figure 1(c) and (d)), vertical (Figure 1(b), (g) and (j)), or tiled (Figure 1(a), (f), and (i)) fashion. Another common method is overlaying (Figure 1(e), (h)). For example, when proportion facts contain geographic information, descriptions are usually overlaid on top of a map.

Layout for Multiple Facts   When combining multiple facts into one, there are four common strategies:

- **Side-by-Side**. To indicate comparison, hierarchy, or other logical relationships, multiple infographics can be arranged in grids, including ones that are parallel, tiled, circular, hierarchical, and stacked (Figure 6(a)).
- **Sharing axes**. This type of layout aligns the visual marks of all single infographics so that they can be placed in a common coordinate system. It is applicable for statistical charts with axes, such as bar charts or scatterplots (Figure 6(b)).
- **Sharing center**. This type of layout arranges multiple numerical facts in the form of concentric circles or sectors with the same center. It is applicable for circular charts, such as pie charts or Nightingale rose charts (Figure 6(c)).
- **Sharing context**. This type of layout links multiple annotations to different positions of shared illustrative images. It is applicable to annotated infographics. Visual elements of multiple info-
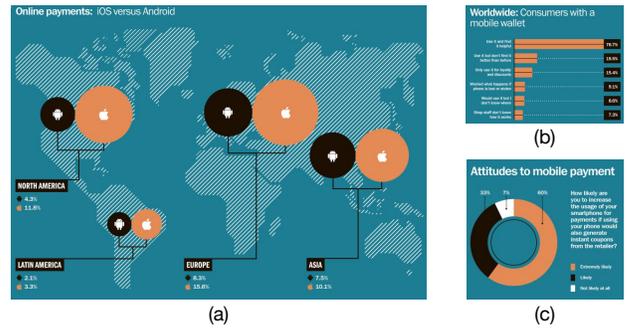

Figure 6. Exemplars of infographics with multiple facts [50]: (a) side-by-side and sharing context, (b) sharing axes, (c) sharing center.

graphics are placed according to the empty space of a common background (Figure 6(a)).

4.2.2 Description

Text description is another key component in data-driven infographics. The original statements provide basic materials of infographic descriptions. Infographic designers usually extract key information and arrange descriptions as text blocks. Each infographic may contain multiple text blocks, combined to deliver a complete message. As mentioned in Section 3, proportion-related facts are about how much a part occupies the whole. Therefore, there are three key pieces of information for each proportion fact: *number*, *part*, and *whole*, which are often explicitly expressed in infographics as descriptions. In addition to these three pieces, we also discovered that designers often treat a number's *modifier* (if there is one), such as "more than" and "less than", as an isolated description. Besides these four key types of descriptions, we also identify the following common forms:

- The entire statement.
- The statement with the *number* removed: For example, "of USA fresh water is used for agriculture" in the statement "40 percent of USA fresh water is used for agriculture."
- The *part* as a verb-object phrase in the statement: Syntactically, when a statement contains a subject and a verb-object phrase, this type of description directly shows the phrase part, for example, "are consumed in breakfast" in the statement "65% of coffee are consumed in breakfast."
- The *number-whole* phrase in the statement: Basically, this type of description shows the subject in a statement, where the *modifier* is optional, for example, "65% of coffee" in "65% of coffee are consumed in breakfast."
- The text segments before and after the *number* in the statement: Based on the position of the *number* component, a statement can also be segmented into three parts and arranged separately, for example, "In the US, less than", "1%", and "men know how to tie a bow tie" in "In the US, less than 1% men know how to tie a bow tie." This segmentation is particularly useful when a statement does not start with the *number* component.

4.2.3 Graphic

The design of graphic components involves the selection and composition of graphics. There are two main considerations for how to display graphics in infographics. First, graphics should be semantically related to the content of the original statements. Second, different graphic elements may have different roles in helping to convey the message.

Based on some existing visualization taxonomies [2, 12], we went through our infographic dataset for proportional facts. We summarized frequently-used visualizations into seven types, namely, pictograph (Figure 1(a)), adornment (Figure 1(d)), donut chart (Figure 1(c)), pie chart (Figure 1 (o)), bar chart (Figure 1(i)), filled icon (Figure 1(e)), and scaled icon (Figure1(g)).

4.2.4 Color

Each infographic has a color theme consisting of harmonious colors, which may also help indicate latent semantics. For example, when an infographic is about environment protection, green or blue based themes are often used. In addition, given a color palette, colors also need to be assigned to different parts of an infographic. The designs in our samples share common rules:

- **Background and foreground**: The background color occupies the biggest area and contrasts the foreground color, while there are often multiple foreground colors used for different visual elements, such as the descriptions, title, and icons. In most cases, descriptions in different text blocks share the same color. Graphic elements, on the other hand, may use one or more colors based on their roles. For example, pictographs, pie charts, or donut charts often require at least two colors to correctly convey the proportion value. Embellishing icons often only use one color to keep a clean look.
- **Number highlight**: Being an important component, *numbers* often require extra emphasis. Most of the samples highlight the values with a distinguished size, color, and font (e.g., Figure 1(a) and (g)). Alternatively, values can be embellished with background pictures (e.g., Figure 1(c) and (i)).

5 TEXT-TO-VIZ IMPLEMENTATION

Our system contains two main modules, namely *text analyzer* and *visual generator*. First, users provide a textual statement about a proportion fact, such as "More than 40% of students like football." Then, our *text analyzer* identifies the essential segments in the statement, including *modifier*, *whole*, *part*, *number*, and others (Section 4.2.2). Then the original statement and the extracted segments are fed into the *visual generator* for infographic construction. For each dimension (i.e., *layout*, *description*, *graphic*, and *color*), a set of visual elements are generated or selected. Then, we enumerate all combinations of these elements, to synthesize valid infographic candidates. Finally, all the synthesized results are evaluated and ranked, and the ones with high scores are recommended to users. Then, users can either directly export anyone of them as an image to integrate into their reports or presentation slides, or select one and further refine it based on their personal preferences.

5.1 Text Analyzer

Given a statement on proportion facts, the goal of this analyzer is to extract the segments of four predefined entity types: modifier (M), number (N), part (P), and whole (W), which is a fundamental task in natural language understanding called named entity recognition. Currently CRF-based approaches have the state-of-the-art performance [29]. In this prototype, for efficiency, we develop a supervised CNN+CRF model to perform the task. Specifically, there are three steps:

- **Tokenization**: First, a sentence is converted into a token sequence through preprocessing, where we take punctuations, abbreviations, and special phrases into consideration. For example, given the statement in Figure 7, we can totally collect a sequence of nine tokens (i.e.,"more", "than", ..., and the final period ".").
- **Featurization**: Then, for each token, we extract multiple features, such as word embedding feature [43], syntactic feature (such as upper/lower case, punctuation, part-of-speech tag), and Brown clustering feature [13]. All these features are concatenated together to form a larger feature vector to represent the token. For the aforementioned example of nine tokens, a $9 \times n$ feature matrix can be obtained, where n (n=2531 in our implementation) is the length of the concatenated feature vector.
- **CNN + CRF**: In this step, we feed the output feature matrix to a one-dimensional Convolutional Neural Network (CNN) [25], on top of which is a Conditional Random Field (CRF) [41] layer. In this step, the CNN layer is used to learn a better feature representation for each token from the input raw feature matrix. For the aforementioned example of nine tokens, our CNN layer will generate a $9 \times m$ feature matrix, where m (m=59 in our implementation) is the number of kernels in the layer. Then the CRF
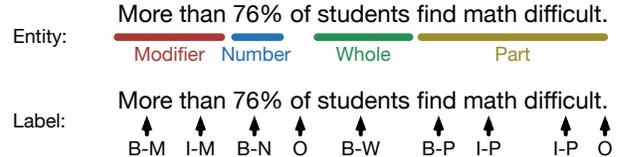

Figure 7. An example of entities and labels in a statement. Following the IOB (inside, outside, and beginning) format [51], we can map the text and entities into a sequence of labels, where B-, I-, and O represent begin, inside, and outside, respectively.

Entity	Modifier	Whole	Part	Number
Precision	1.00	0.81	0.65	1.00
Recall	0.94	0.74	0.73	0.97
F1 Score	0.97	0.77	0.69	0.97

Table 2. Training results for *text analyzer* model.

layer is used to determine the entity labels for all tokens. The parameters of the CNN and CRF are jointly trained.

We train the CNN+CRF model on 800 manually annotated statements. Table 2 shows the performance of 10-fold cross-validation.

5.2 Visual Generator

Taking the information provided by the *text analyzer*, this generator identifies multiple candidates on each design space dimension, including *layout*, *description*, *graphic*, and *color*. After that, all the combinations of these candidates are enumerated to identify valid infographic results in the *synthesis* step. Then, the quality of the resulting infographics is evaluated through the *ranking* step.

5.2.1 Layout

The *layout* module contains a set of blueprints that describe the overall look of the resulting infographics. First, a blueprint specifies the aspect ratio of the resulting infographic and how the space is divided into regions. More importantly, for each region, the blueprint also needs to specify what kind of information should be placed there, along with any constraints concerning the corresponding information.

Figure 8(a) shows an example. This blueprint first specifies that the resulting infographic should have an overall 2×1 aspect ratio. The canvas is then recursively divided into three regions. Although topological relationships are specified, individual regions are allowed to grow or shrink to accommodate their content. Furthermore, we added the following key constraints:

- This blueprint only accepts statements that start with numerical values, e.g., "76% students find math difficult."
- Region 1 holds a graphical element, which can be a pictograph, a filled icon, a donut chart, or a pie chart;
- Region 2 holds the *number* part in the input statement;
- Region 3 holds the the input with the *number* part removed;
- The font size in Region 2 should be at least three times and at most eight times the font size in Region 3.

Clearly, this *layout* module cannot enumerate all creative infographic designs. However, the goal of this system is to provide the quickest way to generate common but professional infographics. Thus, we analyzed the collected samples, identified a set of exemplars, and built 20 layout blueprints in total. These layouts serve as our initial pool of blueprints for infographics generation. However, this is an expandable approach, so we can add more layouts as needed.

5.2.2 Description

The basic descriptions, i.e., *modifier*, *part*, *whole*, and *number*, are identified by the *text analyzer*. To provide additional description candidates (Section 4.2.2) that may be required by different blueprints, we
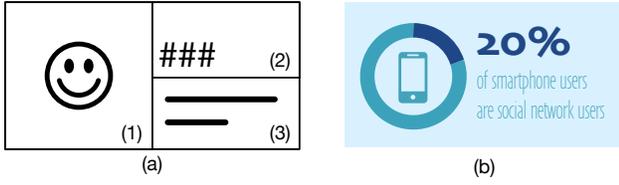
Figure 8. (a) A layout blueprint example and (b) its realization.

adopted the Stanford Parser [20] to analyze the grammatical structure of the input natural language statement. From the generated grammar tree, we extracted descriptions with different lengths and different components to meet the need of different layouts

5.2.3 Graphic

This *graphic* module aims to provide all graphic elements, including representative icons and other data-driven graphics, such as pie charts and donut charts. In addition, various applicable constraints should also be considered here. When integrating icons into layout blueprints, these constraints will decide whether the resulting infographic is a valid one or not. For example, country-shape icons are not suitable for pictographs and are often used as backgrounds (Figure 1(h)) or filled shapes (Figure 1(e)). Icons that represent the *part* are not suitable for pictographs either. Hollow icons also cannot be used as filled shapes, since the colors in them are hardly discernible. In our implementation, we built an icon library with 100 common icons. For each icon, we manually added one or more keywords indicating its semantic meaning, and proper applicable constraints. Then, for non-stop words in the *part* and *whole* components, we use the Word2Vec [43] technique to find the best-matching icons in the icon library.

5.2.4 Color

The *color* module aims to find a set of color palettes for a specific infographic. There are two considerations here. First and foremost, the colors within a palette should be harmonic. In addition, it would be better if the colors match the context of the given statement. For example, if the proportion fact is about environment protection, it is more appropriate for the color theme of the corresponding infographic to contain green or blue colors. Although several techniques [35, 61] have been proposed to identify semantically relevant colors from words, they mainly focus on individual colors and cannot generate harmonic color palettes. Therefore, to address the needs of our *color* module, we adopted a different approach and built a theme library, which contains a set of color palettes. In each palette, various colors are defined with annotations describing their specific uses, such as background, highlights, text, etc. Leveraging the color design knowledge [1, 18], we can ensure that the colors within a palette are harmonic. In addition, we also added one or more keywords to each color palette, describing their preferred context, such as "environment", "fruit", "taxi", and "technology". Then for the non-stop words in the input statements, we again use the Word2Vec [43] technique to find the best-matching palette.

5.2.5 Synthesis

At the *synthesis* step, for each layout blueprint, we enumerate through all recommended graphics, color palettes, and descriptions, and then generate all valid infographics.

First, we rule out all the layout blueprints that require elements nonexistent in the input statement. For example, if a blueprint requires a region to be filled with *modifier* information (e.g., "More than"), but such information is not provided by the *description* module, then this blueprint will be considered invalid and ruled out. Then, for each valid blueprint, we extract the required icon and description information, calculate their aspect rations, and try to put them into the blueprint layout. The goal is to scale them uniformly and maximize their total size, given the predefined layout constraint. We formulate it as a UI

interface optimization problem and solve it using the solver proposed by Badros et al. [6].

Since the final results are optimized based on the aspect ratios of visual elements, we also need to pre-compute several options for pictographs and descriptions. For example, we choose several common pictograph arrangement, such as 2×5, 1×5, and 1×10, to obtain different aspect ratios for the solver to enumerate. Given a description, different line breakings may also yield different aspect ratios. To obtain a set of candidate aspect ratios for a description, we first consider the different line counts (from 1 to 10). Then, we use dynamic programming to obtain a line breaking setting that has the minimum raggedness [31] for each line count, since designers generally prefer similar lengths when a text is broken into multiple lines. In addition, a pre-selected set of fonts with different compactnesses are also enumerated to find the font that yields the best aspect ratio to match the assigned canvas space.

5.2.6 Ranking

Since there may be multiple values recommended by these modules, the number of valid combinations may be large. And clearly they are not equally appealing. Therefore, for each resulting infographic, we need to evaluate its quality of message delivery. In this system, we consider three scores, namely *semantic*, *visual*, and *informative*.

- **Semantic score**: This metric aims to evaluate the quality of the selected icons and themes. Since icons and themes are selected through keyword matching, the better the matching, the higher the semantic score is. For example, given the statement "70% students find math difficult," the icon of "student" should contribute to a higher semantic score than the icon of "person". Thus, we define the semantic score α_s as the average value of all the displayed icons and color palettes picked from the *graphic* and *color* modules using the Word2Vec [43] measure. Since we prefer infographics with graphic elements, we define $\alpha_s = 0$ if there are no matching icons or color palettes.

- **Visual score**: This metric aims to evaluate the visual appearance of an infographic. Since the aspect ratio of icons or the length of the description embedded in the infographic may not be ideal to the layout design, icons and text may be scaled to fit into the layout. Thus, this metric is designed to measure the empty space that these visual elements waste. The better they fit, the higher the visual score is. Thus, we define the visual score as:

$$\alpha_v = \frac{\text{the area of all displayed elements}}{\text{the area of the canvas}}$$

- **Informative score**: Ideally, readers should easily recover the original information from the resulting infographic. However, due to the applicable constraints of these values, some information may be left out in infographics, which clearly should be avoided. For example, shorter descriptions may be used due to space limitations. Therefore, we propose the informative metric to evaluate the completeness of message delivery. Thus, we define the informative score as $\alpha_i = \sum_{\omega \in S} I(\omega)/|S|$, where S denotes all the non-stop words in the input statement, and

$$I(\omega) = \begin{cases} 1 & \text{if } \omega \text{ appears in the result as an icon or word} \\ 0 & \text{otherwise.} \end{cases}$$

The total score is defined as a weighted sum of all three scores: $\alpha = w_s \alpha_s + w_v \alpha_v + w_i \alpha_i$. The default values of these weights are set to 0.25, 0.5, and 0.25, respectively, based on our experiments.

5.2.7 User Refinement

Although we realize that different users may have different tastes and aim to provide a variety of choices so that users can pick one that best meets their personal tastes, it is still possible that users need to perform minor touches to the final results. Without compromising too much of the intuitiveness of our system, we decided to support users replacing elements in each infographic, including icons, color palettes, and descriptions. By clicking on a recommended infographic, users can open

a dialog, in which all the adopted icons, colors, and descriptions are listed. Then users can easily replace them with whatever alternatives they like. When users are satisfied with the result, they can easily save the template or infographic for the reuse purpose.

6 EVALUATION

In this section, we first present a set of examples to demonstrate the diverse designs that our system can automatically create. To understand how general users perceive our system, we further conducted a set of casual interviews with a wide variety of audiences in two exhibit events. The last assessment involved expert evaluation with three professional graphic designers.

6.1 Sample Infographics

To demonstrate the capabilities of Text-to-Viz, we present a variety of infographics created with our system (Figure 1). For example, Figure 1(a)-(d) and Figure 8(b) are all generated based on the same statement, "More than 20% of smartphone users are social network users." We can see that different templates can produce different infographics. In particular, we can see that Figure 8(b) is based on the layout blueprint illustrated in Figure 8(a). However, since the template does not reserve a place for the *modifier* component, the generated infographic is less accurate than the others, and hence has a lower informative score. Figure 1(e) shows an example of a filled icon, while Figure 1(f) shows an example of a tilted layout. Figure 1(g) and (h) demonstrate two examples of how our system handles proportion information in the form of "*m in n*". Our system can either choose the correct number of icons to form a pictograph (Figure 1(g)) or convert the information to a percentage number and show it with other visualizations (Figure 1(h)). Figure 1(i) and (j) show that our color strategy can correctly select colors based on semantic information. Since one of our color palettes has the descriptive keyword *coffee*, this color palette will be ranked higher when choosing colors for infographics. Figure 1(k)-(o) demonstrate the results for showing multiple percentages. Specifically, Figure 1(k)-(m) show a *comparison* case, in which proportion facts cannot be logically accumulated, while Figure 1(n) and (o) show an *accumulation* case.

6.2 Casual User Interview in Exhibits

Our system has been demonstrated to the public in two exhibits. One was a half-day exhibit with an audience consisting mostly of journalists, and the other event was a two-day exhibit with an audience consisting of employees from all departments of a tech company, including sales, software developers, program managers, and public relation managers, etc. In the two events, we received more than 80 different audience members in total. During each reception, we first introduced the background and system and gave them a short live demo to illustrate the idea. Then, we encouraged them to explore the system using their own statements. During the discussions, we focused on the following questions:

- Do you think this tool is useful?
- Do you like the results generated?
- Do you have any suggestions considering your background?

Overall, the feedback is overwhelmingly positive. When we explained the problem to the audience members, they understood immediately, since they had all encountered similar situations as described in Section 1. Since most of them did not have a design background, this tool would provide much greater convenience when they need infographics. Several audience members described all the troubles they had gone through to build an infographic and how they had finally given up after several attempts, as they felt the tools were too complicated for them. Given their different backgrounds, the audience members also suggested a variety of applications for this tool. The journalists immediately saw the value of it for their reports. Several public relation managers and project managers believed this tool would greatly help their presentations. In particular, one development manager asked if we could support speech-to-infographics and said, "It would be very interesting to see infographics popping up on screens when people are discussing in a meeting."

One compromise we made for this system was creativity. Since our infographic layouts are based on templates, we were worried that the output infographics may also lack creativity and feel homogeneous. However, when we asked the audience how they felt about the output results, to our surprise, none of them unpromptedly noticed the homogeneity and creativity issue. All of them commented that the outputs were very impressive and professional and that they would love to directly use them out of the box. We suspected that this may be because they, again, lacked design experiences and only used our system momentary. However, we did encounter some problems when we asked the users to try our system with their own statements. The most critical problem was the recommended icons, which sometimes did not match the expectations from users. For example, when a lab director input the statement: "30% secretaries wear glasses", the recommended icon was a woman's skirt, which seemed inappropriate. This was caused by the Word2Vec matching algorithm. We also noticed that the audience members were the most sensitive to icons, which is probably because icons are more interpretable by non-designers, compared to other visual channels, such as *layout*, and *color*. Therefore, we should put more effort into generating meaningful icons from text in the future, to ensure more satisfactory infographics overall. Meanwhile, we allow users to manually replace icons as a temporary solution.

The most frequently asked question was when we would make this tool publicly available. In addition, the audience also asked if we could support more types of statements. Both requirements are expected since they can certainly help the audience members with their tasks. On the other hand, we also collected some unexpected feedback, which indicated us a future direction for this project. For example, one audience member asked if our algorithm could automatically adjust colors or styles to make it match an existing document or PowerPoint presentation. Another data scientist suggested that we could perhaps generate infographics directly from a dataset. By connecting to a data analysis module, our system can become more intelligent and then be used in many other scenarios.

6.3 Expert Interview

To understand the effectiveness of our system, we conducted a user study with three designers in a tech company. All of them had graduated (E1, E3) from or were enrolled (E2) in professional schools of design disciplines. E1 majored in user experience design and had more than four years of design experience; E2 was a student majoring in industrial design with more than two years of design experience; and E3 majored in visual communications and had more than eight years of experience in design and two years of experience in photography. Their current roles are user interface designer and design intern.

The 60-minute study started with a 10-minute introduction of our tool, including the usages, several use cases, and the sample infographic designs it generated. After that, a semi-structured interview was conducted to understand the designers' views on the workflow, the tool's usability, and the resulting infographics.

Overall, the designers agreed that the tool has very promising usages. They felt that the results generated are very impressive and can be used for different purposes, such as posters, presentation slides, and data reports. All of them said they would like to use it when available.

Our participants appreciated the overall design of the workflow. They expressed the needs for using such a system in their daily work and felt that it would be very convenient to have text as input and a list of candidate designs as output. "It can be used in many productivity tools to enhance data presentation," said one of the designers. Further, E1 suggested that it should be integrated into daily work settings, which would improve the authoring process to be transparent and without interruption. E1 also mentioned, "If it is not well-integrated, I would like to have a copy-to-clipboard button to reduce the effort of using it." In terms of the authoring functions we have provided, such as changing icons and color themes, the experts felt that there is no need to incorporate more editing, saying, "It is obvious that the target of this tool is not for producing very creative and elaborate designs, so it is great to keep it simple and clear."

In terms of the quality of the generated infographics, the designers

thought it was good enough considering its target user group. E1 said, "Users may be lazy to seek perfection. A relatively good design using little effort is sufficient in most cases." For most of the use cases, the designers felt that they were flexible and diverse enough. They even commented: "The number of choices is *not the more the better*'. Too many choices may become a burden and make it hard to make decisions." They also gave advices on the resulting infographics. For example, E1 said, "Some of the infographics involve more than two fonts, but I think using one is ideal." E3 thought it would be acceptable to sacrifice visualization accuracy for better visual appealingness. For example, for the number "65%", E3 expressed that it would be better to fill in two out of three icons. They also made suggestions on further considering the semantic meaning of the icons. For example, when filling a cup or waterdrop icon with colors, it should be bottom-to-top instead of left-to-right.

7 DISCUSSION

7.1 Automatic Generation of Infographics

Traditionally, infographic design is considered a highly creative task that can hardly be performed solely by computers. Therefore, although many interactive tools have been developed to build infographics, they still maintain a manual editing process, in which designers lead and computers assist. This is not a suitable paradigm for casual users, who are also a major category of infographic users overlooked by previous research. Unlike designers, casual users neither possess design expertise nor are they familiar with advanced tools, which has also been repeatedly confirmed during our casual interviews (Section 6.2). Therefore, when they need an infographic in their presentations or reports, they must either go through the troubles of creating one by themselves, or ask a professional designer to make one for them. Neither approach seems to have a good return-of-investment, especially when their expectations for the infographic are not high. In our survey, we indeed found that, despite creative designs, there are simpler designs that are used repeatedly in various samples, which leads us to believe that many people (even designers) still favor simple and clear infographics, and do not like overly-embellished designs. This scenario motivated our work: to generate sufficiently good infographics in a most natural way, using natural language statements, for casual users.

7.2 Opportunities for New Usage

The opportunities for this solution are enormous. With little or no human interactions, infographics can now be produced effortlessly, hence attracting more and more casual users to take advantage of infographics in their daily life. However, our proof-of-concept implementation may not be the best use scenario for this kind of technology, since it is still an interruptive experience. Users have to stop their work at hand, open our system, enter the statement, then obtain a proper result to insert back into their work. However, we anticipate that this tool can work seamlessly with productivity software suites, such as Office and iWork. When users are working on their presentations and reports, the text analysis module can run silently in the background. Once it detects a statement that can be augmented with infographics, a message would come prompt up to check with users whether they want to take a recommended infographic. In this way, infographics can reach the broadest range of audience and be helpful to them.

Another interesting opportunity is poster authoring. An infographic poster often contains multiple infographics that are semantically related. Although each infographic can be generated independently, combining them as a poster can deliver a stronger or more complex message. To achieve this goal, additional considerations, such as typesetting and visual consistency, should also be put into this equation.

In this paper, we demonstrate a proof-of-concept system that takes natural language statements as input. However, we think it is possible to connect the system to other data sources, such as database or documents. Although we may need to incorporate more machine learning techniques to extract interesting patterns first, it will also greatly expand the applicability of our approach, e.g., integrated into professional visual analytics systems for data analysis tasks.

7.3 Failure Cases and Limitations

During the development of our prototype, we also observed some failure cases, in which our system generates wrong or bad results. One critical source of these failure cases is the *text analyzer*. For some complicated and long statements, such as "funds for administration were limited to 15% of the scholarship amount," the analyzer may not segment correctly and provide correct tags, which will eventually lead to incomprehensible infographics. We believe more training data can mitigate this situation. In practical use, we can also increase the prediction threshold to reduce the number of inaccurate predictions to surface. Another main source is our icon selector. Many bad cases we observed are caused by inappropriately matched icons. In addition, the length and wrapping of descriptions may also cause less pleasing or readable infographics. For example, if a description is too long, our system will automatically select a small font size, which may result in an inharmonious or even illegible infographic. Although we have integrated a ranking mechanism to help evaluate the overall quality of generated infographics, there are still many aspects of aesthetic to consider in the future.

In addition, our design also has some limitations in terms of capability. The first and foremost limitation is that our current approach can only handle a relatively small set of information. However, there are various types of information that can be represented by infographics. So far, our approach has to manually identify key information and visual representations type by type. Although in this paper, we demonstrate the auto-generation paradigm on one type of information (proportion facts), it is still unknown how this paradigm extends to other types of information. The second limitation is infographic expressiveness. Clearly, this approach lacks human creativity and is based on a set of pre-designed infographic styles. Although it is an open framework that allows users to add more materials to enrich the infographic designs, the resulting infographics are still limited. The third limitation concerns expression ambiguity. For example, consider these two statements: "30% of students are French; while 40% are American," and "30% of students speak French, while 40% speak English." We can see that in the first statement, the two percentages can be aggregated, which makes it is acceptable to only use one pie chart to visualize the information. However, for the second statement, it is not appropriate to combine the two values in one pie chart, since there may be students who speak both French and English. This kind of ambiguity is extremely hard to resolve using our current machine learning model. Thus, a model that incorporates a deeper understanding of knowledge is required. What adds to the problem is user intent. Even for the former expression, users may still choose one or two pies to emphasize the aggregation or comparison, although both are reasonable visuals. The user intent is impossible to infer from the above examples. In our implementation, we do not tackle this ambiguity problem and leave the decision to users.

8 CONCLUSION AND FUTURE WORK

In this paper, we introduce a framework to automatically generate infographics and demonstrate its feasibility through a proof-of-concept system. Our system takes a natural language statement on a proportion fact and translates it to a set of professional infographics with different styles and settings, which casual users select from or refine based on their personal preferences. Our approach does not require a complex authoring process or design expertise. The example results and user/designer interviews show the tool's usability and promise of being adopted in everyday life.

There are several avenues that are promising for future work. As a proof-of-concept, our system works relatively well and is only for a specific type of information. In the future, we would like to expand our work to support more types of statistical information or even other types of infographic, such as timelines and locations. On the other hand, the implementation of our current system is also limited. For example, many algorithms used here are still rule-based, such as icon selection, color selection, and ranking. We believe it will be very interesting to explore more techniques (especially machine learning based techniques) to further improve the result quality of our current system.

REFERENCES

[1] Adobe. Adobe color cc. `https://color.adobe.com`.

[2] F. Amini, N. H. Riche, B. Lee, A. Monroy-Hernandez, and P. Irani. Authoring data-driven videos with dataclips. *IEEE transactions on visualization and computer graphics*, 23(1):501–510, 2017.

[3] M. Artacho-Ramirez, J. Diego-Mas, and J. Alcaide-Marzal. Influence of the mode of graphical representation on the perception of product aesthetic and emotional features: An exploratory study. *International Journal of Industrial Ergonomics*, 38(11-12):942–952, 2008.

[4] M. Asahara and Y. Matsumoto. Japanese named entity extraction with redundant morphological analysis. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 8–15. Association for Computational Linguistics, 2003.

[5] B. Bach, Z. Wang, M. Farinella, D. Murray-Rust, and N. Henry Riche. Design patterns for data comics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 38. ACM, 2018.

[6] G. J. Badros, A. Borning, and P. J. Stuckey. The cassowary linear arithmetic constraint solving algorithm. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 8(4):267–306, 2001.

[7] S. Bateman, R. L. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks. Useful junk?: the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2573–2582. ACM, 2010.

[8] J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, 2013.

[9] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel. Nymble: a high-performance learning name-finder. In *Proceedings of the fifth conference on Applied natural language processing*, pages 194–201. Association for Computational Linguistics, 1997.

[10] R. Borgo, A. Abdul-Rahman, F. Mohamed, P. W. Grant, I. Reppa, L. Floridi, and M. Chen. An empirical study on using visual embellishments in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2759–2768, 2012.

[11] 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, 2016.

[12] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2306–2315, 2013.

[13] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

[14] C. Bryan, K.-L. Ma, and J. Woodring. Temporal summary images: An approach to narrative visualization via interactive annotation generation and placement. *IEEE transactions on visualization and computer graphics*, 23(1):511–520, 2017.

[15] Z. Bylinskii, S. Alsheikh, S. Madan, A. Recasens, K. Zhong, H. Pfister, F. Durand, and A. Oliva. Understanding infographics through textual and visual tag prediction. *arXiv preprint arXiv:1709.09215*, 2017.

[16] L. Byrne, D. Angus, and J. Wiles. Acquired codes of meaning in data visualization and infographics: beyond perceptual primitives. *IEEE transactions on visualization and computer graphics*, 22(1):509–518, 2016.

[17] I. Cash. Infographic of infographics. `http://www.ivan.cash/infographic-of-infographics`.

[18] Coolors. coolors. `https://coolors.co/browser`.

[19] Z. Cui, S. K. Badam, M. A. Yalçin, and N. Elmqvist. Datasite: Proactive visual data exploration with computation of insight-based recommendations. *Information Visualization*, 18(2):251–267, 2019.

[20] M.-C. De Marneffe, B. MacCartney, C. D. Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of Language Resources and Evaluation Conference*, pages 449–454. Genoa Italy, 2006.

[21] Ç. Demiralp, P. J. Haas, S. Parthasarathy, and T. Pedapati. Foresight: Recommending visual insights. *Proceedings of the VLDB Endowment*, 10(12):1937–1940, 2017.

[22] T. Gao, M. Dontcheva, E. Adar, Z. Liu, and K. G. Karahalios. Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pages 489–500. ACM, 2015.

[23] S. Haroz, R. Kosara, and S. L. Franconeri. Isotype visualization: Working memory, performance, and engagement with pictographs. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 1191–1200. ACM, 2015.

[24] L. Harrison, K. Reinecke, and R. Chang. Infographic aesthetics: Designing for the first impression. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1187–1190. ACM, 2015.

[25] B. Hu, Z. Lu, H. Li, and Q. Chen. Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems*, pages 2042–2050, 2014.

[26] K. Z. Hu, M. A. Bakker, S. Li, T. Kraska, and C. A. Hidalgo. Vizml: A machine learning approach to visualization recommendation. *arXiv preprint arXiv:1808.04819*, 2018.

[27] J. Hullman, E. Adar, and P. Shah. The impact of social information on visual judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1461–1470. ACM, 2011.

[28] J. Hullman, N. Diakopoulos, and E. Adar. Contextifier: automatic generation of annotated stock visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2707–2716. ACM, 2013.

[29] M. Ju, M. Miwa, and S. Ananiadou. A neural layered model for nested named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 1446–1459, 2018.

[30] N. W. Kim, E. Schweickart, Z. Liu, M. Dontcheva, W. Li, J. Popovic, and H. Pfister. Data-driven guides: Supporting expressive design for information graphics. *IEEE transactions on visualization and computer graphics*, 23(1):491–500, 2017.

[31] D. E. Knuth and M. F. Plass. Breaking paragraphs into lines. *Software: Practice and Experience*, 11(11):1119–1184, 1981.

[32] H.-K. Kong, Z. Liu, and K. Karahalios. Internal and external visual cue preferences for visualizations in presentations. In *Computer Graphics Forum*, volume 36, pages 515–525. Wiley Online Library, 2017.

[33] R. Kosara. Presentation-oriented visualization techniques. *IEEE computer graphics and applications*, 36(1):80–85, 2016.

[34] M. Lamm, A. T. Chaganty, C. D. Manning, D. Jurafsky, and P. Liang. Textual analogy parsing: What's shared and what's compared among analogous facts. *arXiv preprint arXiv:1809.02700*, 2018.

[35] S. Lin, J. Fortuna, C. Kulkarni, M. Stone, and J. Heer. Selecting semantically-resonant colors for data visualization. In *Computer Graphics Forum*, volume 32, pages 401–410. Wiley Online Library, 2013.

[36] Z. Liu, J. Thompson, A. Wilson, M. Dontcheva, J. Delorey, S. Grigg, B. Kerr, and J. Stasko. Data illustrator: Augmenting vector design tools with lazy data binding for expressive visualization authoring. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 123. ACM, 2018.

[37] Y. Luo, X. Qin, N. Tang, and G. Li. Deepeye: Towards automatic

data visualization. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 101–112. IEEE, 2018.

[38] J. Mackinlay, P. Hanrahan, and C. Stolte. Show Me: Automatic presentation for visual analysis. *IEEE transactions on visualization and computer graphics*, 13(6):1137–1144, 2007.

[39] S. Madan, Z. Bylinskii, M. Tancik, A. Recasens, K. Zhong, S. Alsheikh, H. Pfister, A. Oliva, and F. Durand. Synthetically trained icon proposals for parsing and summarizing infographics. *arXiv preprint arXiv:1807.10441*, 2018.

[40] L. Màrquez and H. Rodríguez. Part-of-speech tagging using decision trees. In *European Conference on Machine Learning*, pages 25–36. Springer, 1998.

[41] A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 188–191. Association for Computational Linguistics, 2003.

[42] G. G. Méndez, M. A. Nacenta, and S. Vandenheste. iVoLVER: Interactive visual language for visualization extraction and reconstruction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4073–4085. ACM, 2016.

[43] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[44] B. Mitchell. Behind the internet curtain. https://www.digitalrealty.com/blog/behind-the-internet-curtain, 2014.

[45] A. V. Moere and H. Purchase. On the role of design in information visualization. *Information Visualization*, 10(4):356–371, 2011.

[46] A. V. Moere, M. Tomitsch, C. Wimmer, B. Christoph, and T. Grechenig. Evaluating the effect of style in information visualization. *IEEE transactions on visualization and computer graphics*, 18(12):2739–2748, 2012.

[47] D. Moritz, C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe, and J. Heer. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE transactions on visualization and computer graphics*, 25(1):438–448, 2019.

[48] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[49] H. C. Purchase, K. Isaacs, T. Bueti, B. Hastings, A. Kassam, A. Kim, and S. van Hoesen. A classification of infographics. In *International Conference on Theory and Application of Diagrams*, pages 210–218. Springer, 2018.

[50] Raconteur. Mobile payments world view. https://michaelrosensays.wordpress.com/tag/giving-usa-2015-infographic, 2015.

[51] L. A. Ramshaw and M. P. Marcus. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer, 1999.

[52] L. F. Rau. Extracting company names from text. In *Artificial Intelligence Applications, 1991. Proceedings., Seventh IEEE Conference on*, volume 1, pages 29–32. IEEE, 1991.

[53] 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)*, pages 230–239. IEEE, 2017.

[54] D. Ren, T. Höllerer, and X. Yuan. iVisDesigner: Expressive interactive design of information visualizations. *IEEE transactions on visualization and computer graphics*, 20(12):2092–2101, 2014.

[55] D. Ren, B. Lee, and M. Brehmer. Charticulator: Interactive construction of bespoke chart layouts. *IEEE transactions on visualization and computer graphics*, 25(1):789–799, 2019.

[56] F. Richter. Ibm tops u.s. patent ranking for 21st consecutive year. https://www.statista.com/chart/1796/us-patent-ranking-2013, 2014.

[57] M. J. Rosen. Strong american philanthropy at a record high! https://michaelrosensays.wordpress.com/tag/giving-usa-2015-infographic, 2015.

[58] A. Satyanarayan and J. Heer. Lyra: An interactive visualization design environment. In *Computer Graphics Forum*, volume 33, pages 351–360. Wiley Online Library, 2014.

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

[60] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang. Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 365–377. ACM, 2016.

[61] V. Setlur and M. C. Stone. A linguistic approach to categorical color assignment for data visualization. *IEEE transactions on visualization and computer graphics*, 22(1):698–707, 2016.

[62] W. V. Siricharoen. Infographics: the new communication tools in digital age. In *The international conference on e-technologies and business on the web (ebw2013)*, pages 169–174. The Society of Digital Information and Wireless Communication, 2013.

[63] D. Skau and R. Kosara. Readability and precision in pictorial bar charts. In *Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Short Papers*, pages 91–95. Eurographics Association, 2017.

[64] 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.

[65] J. Stegman. 3 way to grow your support revenue. https://www.tsia.com/blog/infographic-3-ways-to-grow-your-support-revenue, 2015.

[66] Y. Sun, J. Leigh, A. Johnson, and S. Lee. Articulate: A semi-automated model for translating natural language queries into meaningful visualizations. In *International Symposium on Smart Graphics*, pages 184–195. Springer, 2010.

[67] E. R. Tufte. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 2001.

[68] F. B. Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon. Manyeyes: a site for visualization at internet scale. *IEEE transactions on visualization and computer graphics*, 13(6), 2007.

[69] Y. Wang, H. Zhang, H. Huang, X. Chen, Q. Yin, Z. Hou, D. Zhang, Q. Luo, and H. Qu. Infonice: Easy creation of information graphics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 335. ACM, 2018.

[70] T. Willingham. Alligator pear: Imports and exports. https://www.freightwaves.com/news/infographics/alligator-pear-import-export, 2019.

[71] T. Winograd. Understanding natural language. *Cognitive psychology*, 3(1):1–191, 1972.

[72] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics*, 22(1):649–658, 2016.

[73] H. Xia, N. Henry Riche, F. Chevalier, B. De Araujo, and D. Wigdor. Dataink: Direct and creative data-oriented drawing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 223. ACM, 2018.