An Ontological Approach towards Automatic Creation of Infographics from Formal Text (Student Abstract)

Devin Garg, Tanuj Agarwal and **Chiranjoy Chattopadhyay**

Indian Institute of Technology Jodhpur, India
{garg.7,agarwal.5,chiranjoy}@iitj.ac.in

Abstract

Infographics deal with representing data or information visually in a perceptually compelling manner. Recently, infographics have gained widespread popularity, giving rise to automated infographics synthesis from texts. Our research follows an ontological approach to automatically extract the necessary indicators from an input sentence and synthesize an infographic corresponding to it. This work includes (1) the creation of a dataset, (2) an end-to-end domain-agnostic framework, and (3) demonstrating the application of the proposed framework. The results demonstrate our framework's ability to extract the necessary textual cues from real-world textual descriptions (from various domains) and synthesize meaningful infographics.

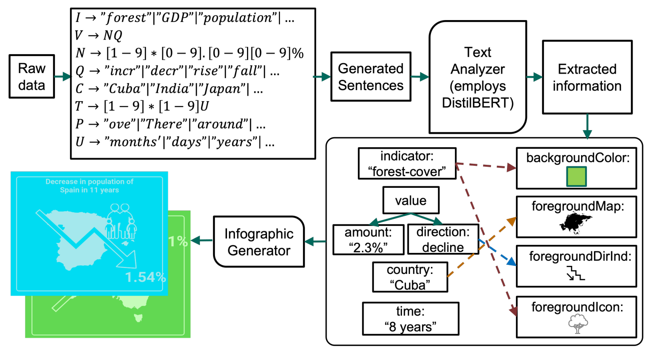


Figure 1: An illustration of the proposed framework.

Introduction

In this age of an overwhelming amount of information and data around us, it is vital to make sense of it. With the widespread use of the internet and technological improvement, the reach of data is more than ever before. To this effect, one needs to identify the 'correct' means for disseminating and understanding this data. It can be in numerous forms, including (but not limited to) textual, visual, audio-visual, etc. Research has shown that visual content in the form of infographics, charts, etc., are associated with improved memory and engagement as opposed to plain original data (Haroz, Kosara, and Franconeri 2015; Bateman et al. 2010). Informational graphics or 'infographics' are units of data representation that contain visual elements and data to make it more presentable and cognitively easy to consume.

The process of creating infographics is domain-specific and hence is time and resource-consuming in nature. As the technology progresses, the need for automatic infographic generation has gained much interest. With a domain-agnostic process at hand, it is possible to synthesize an infographic automatically. These can then be used directly or post further improvement by domain experts as per their requirements. In the literature, there have been attempts to extract information from natural language text with the help of neural networks. This 'useful' information is then used to create the graphics either by using some predefined templates (Cui et al. 2019) or by scraping design templates from infographics on the internet and then fitting the information into these templates, fine-tuning them in the process (Qian et al. 2020) to generate the final product.

Domain Agnostic Ontological Framework

Two constraints came to light during the literature survey, (1) the lack of any open-source dataset, (2) the variability in natural language that influenced the researchers to tackle statements categorically for this task. In this research, we propose a 3-step process for automatic infographic generation. The input expected by the framework is a sentence in the English language, and the output is an infographic. Firstly, following the current approach, the input sentence is passed through a text analyzer. We have adapted DistilBERT (Sanh et al. 2019) pre-trained on the task of named entity recognition (NER). However, this choice is not absolute, and it can be replaced by another model suited to the NER task. The reason behind using DistilBERT was that it is a lightweight model that any researcher can quickly deploy. This text analyzer extracts the labels that are central to the sentence's domain. These labels are then sent to the second part of the framework that generates a mapping from the labels to the graphical elements in the final infographic using a resource description framework (RDF) schema. This schema is used for ontological representation.

An RDF scheme has some classes which need to be defined as per the domain of the sentence. The *output classes* represent various elements of the infographic. As an example, for the sentence: *Over the past **8 years**, **Bhutan** has experienced a jump of around **2.58%** in its **forest cover**.*, the classes concerning the domain would be the period, the country, the indicator and the change in the indicator value. These classes are then mapped to the *output classes* which could be a background image, foreground image, etc.

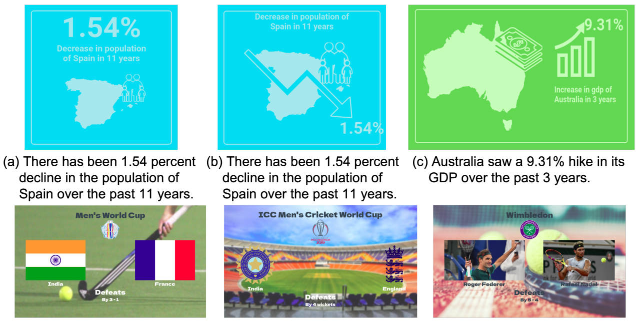
As shown in Figure 1, the classes in the left are domain-dependent and need to be put in place by the user. Such a mechanism is carried out in conjunction with the corresponding mapping to the output graphic's classes shown on the right side of the ontological representation. Such a representation paved the way to a formal framework for mapping extracted information from sentences to elements of the output infographic. Finally, the mapping indicated by the ontology is built with the help of dictionaries of the content related to the domain.

Grammar for Dataset Generation

To tackle the challenge of datasets, we developed a grammar to generate sentences that researchers can use to train the text analyzer module (mentioned in the above section) on the task of named entity recognition. We have curated a dataset spanning several domains. For example, we captured values of several indicators for countries from the World Bank Open Data website, having indicators like forest cover, GDP, population, per capita electric power consumption. We scraped data for the past 16 years from the website for all countries for which it was available. Using that data, we generated natural language statements that represented a change in the value of an indicator for a country over a given period. To create statements in a structured and coherent manner across domains, we propose a grammar consisting of parameters representing quantities that are the essential helpful information that we intend to extract. This 'useful' information is interspersed with placeholder words and then expressed in an infographic as an end goal of the process.

Results and Generalization

In the beginning, we considered proportion-based sentences following previous research (Cui et al. 2019), (Qian et al. 2020). However, since the framework we developed is generalized, we considered a completely different domain for testing the approach. We used sports results as the raw text. Using such sentences, we generated infographics by suitably substituting the appropriate domain-specific components in the ontology. For instance, *Djokovic thrashes Nadal by* $3-0$ *to clinch the Australian Open*. This kind of sentence would have 'useful' information like player names, title, and score extracted from it. We have used this information to create an infographic using the mapping defined in the ontology. Some results obtained for the two domains are shown in Figure 2. The top row depicts proportion-based infographics, while the bottom row depicts the sports domain. In the top row, for the same caption, two variants of infographics have been generated, while in the bottom row, to indicate the country, either flag or the logo of the board has been used. These instances reveal the diversity that the proposed framework allows while designing an infographic.



(a) There has been 1.54 percent decline in the population of Spain over the past 11 years. (b) There has been 1.54 percent decline in the population of Spain over the past 11 years. (c) Australia saw a 9.31% hike in its GDP over the past 3 years.

(e) India defeat France, 3-1 and win the Men's Hockey World Cup. (e) India beat world no 2 England by 4 wickets to clinch the ICC Men's Cricket World Cup. (f) Federer beats Nadal, 5-4, in a nail-biter to emerge as the Wimbledon champion.

Figure 2: Qualitative results of infographics synthesis. The caption of the sub-figures is the input sentences corresponding to the respective infographics.

For someone to use this framework for a new domain, the following steps are: first, fine-tune the text analyzer specifically to the targeted field. Our experimental study reveals that a few sentences (800 to be precise) were required to learn the structure that the sentences of the domain follow. We have data augmentation, which involves back-translation, random deletion, synonym replacement, and noise insertion. The second step would be defining the ontological relationship that is needed between the extracted information and the graphic's elements. Finally, one must specify the sources for obtaining the required visual assets, such as icon libraries and image dictionaries.

Conclusions and Future Work

The way this task of infographic generation has been approached until now has been mostly domain-centric. In this work, however, we have proposed an end-to-end framework for the generalized task. The proposed steps would improve upon the intermediate stage by using a learned ontological mapping to reduce user intervention.

References

Bateman, S.; Mandryk, R. L.; Gutwin, C.; Genest, A.; McDine, D.; and Brooks, C. 2010. Useful junk? The effects of visual embellishment on comprehension and memorability of charts. In *ACM CHI*.

Cui, W.; Zhang, X.; Wang, Y.; Huang, H.; Chen, B.; Fang, L.; Zhang, H.; Lou, J.-G.; and Zhang, D. 2019. Text-to-viz: Automatic generation of infographics from proportion-related natural language statements. *IEEE T-VG*.

Haroz, S.; Kosara, R.; and Franconeri, S. L. 2015. Isotype visualization: Working memory, performance, and engagement with pictographs. In *ACM CHI*.

Qian, C.; Sun, S.; Cui, W.; Lou, J.-G.; Zhang, H.; and Zhang, D. 2020. Retrieve-Then-Adapt: Example-based Automatic Generation for Proportion-related Infographics. *IEEE T-VG*.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.