Towards Visualisation Specifications from Multilingual Natural Language Queries using Large Language Models

Maeve Hutchinson[1*], Aidan Slingsby[1], Radu Jianu[1], Pranava Madhyastha[1*]

[1]Department of Computer Science, City, University of London
[*] Corresponding Authors

Abstract

In this paper, we present an empirical demonstration of a prompt-based learning approach, which utilizes pre-trained Large Language Models to generate visualization specifications from user queries expressed in natural language. We showcase the approach's flexibility in generating valid specifications in languages other than English (e.g., Spanish) despite lacking access to any training samples. Our findings represent the first steps towards the development of multilingual interfaces for data visualization that transcend English-centric systems, making them more accessible to a wider range of users.

CCS Concepts

• ***Computing methodologies*** → *Natural language processing;* • ***Human-centered computing*** → *Visualization systems and tools;*

1. Introduction

Current systems for creating data visualisations require users to have the necessary knowledge to use them, either through code or interaction with a GUI. This restricts bespoke visualisation and analysis to those who are able to access these tools. Generating visualisation from natural language (NL2VIS) is an increasingly feasible technique [WWS*21; SS17; NSS21; LTL*22; LHJY21]. Most such systems use either rule-based or deep learning-based methods, which can struggle to handle underspecified or ambiguous queries [TS19], and are only able to take English language queries. Recently, there has been significant research into the development of pre-trained Large Language Models (LLMs). These may be able to address shortcomings of previous approaches as they are pre-trained on large amounts of human language data, often from multiple languages, and have been shown to have necessary inductive biases to go beyond surface level understanding of lexical content. This has allowed for the development systems that better understand user intent [WTB*22]. Some NL2VIS systems have explored the use of LLM-based methods with some success [MS23; Dib23] but have predominantly focused on English language.

We posit that a multilingual tool would allow diverse group of users to express their goals in their native languages more effectively and efficiently. The long-term goal through such multilingual interfaces would be to create an accessible, multilingual Natural Language Interface (NLI) that engages with users through human-computer dialogues, inferring their intent, thus aiding in iterative visual analysis. In this paper, we present the first steps towards this goal. We demonstrate that LLMs can be used to generate Vega-Lite specifications from a natural language query in both English and Spanish languages and showcase the viability of such an NLI.

2. Experimental Setup

Dataset. We use the the corpus by SRINIVASAN, NYAPATHY, LEE, et al. [SNL*21] which consists of 814 natural language utterances, each corresponding to a specific Vega-Lite visualisation. The dataset was collected in an online study where participants were shown both a dataset and a corresponding visualisation, and asked to enter a natural language query that they would hypothetically enter into an NLI to specify the given visualisation. Ten specific charts (histogram, bar, scatter, and line) were created from three datasets.

Model. BLOOM is a Transformer [VSP*17] based autoregressive language model which models the probability distribution over sequences of tokens (where, tokens could be either lexical elements or units of programming code) [WSF*23]. BLOOM is openly available, with comparable performance to state-of-the-art language models. BLOOM was built over carefully curated data which includes data in multiple languages and large GitHub-based code repositories, in contrast to language models (trained only on language data) and closed language models (where training data is not revealed, e.g. GPT-3 [BMR*20]). We present results with the BLOOM 560-million parameter version.

Prompt Engineering. Our central experimental methodology

DOI: 10.2312/evp.20231072

draws inspiration from the prompt-based learning approach, which has proven to be highly effective in numerous Natural Language Processing (NLP) tasks [BMR*20]. The approach involves engineering a natural language prompt input that contains a task description and a set of valid in-context examples. A single in-context example in the prompt is also known as one-shot learning.

In our experiments, we categorise in-context examples further as a) in-domain; and b) out-of domain examples. An in-domain example is one that contains the solution to a *query* that is semantically similar to the user query within the same dataset. We note that, in our setup the solution corresponds to a fully specified visualization specification. An out-of-domain example is one that is not associated with the same dataset as the user query. We aim to investigate the difference in performance when the model has an input that is structurally similar to the query compared to receiving unrelated information. For the multilingual setup, we present a pilot study where each user query (and user query alone) is machine translated into Spanish language using the OPUS-MT model [TT20] and paired with an in-domain example in the English Language. Thus, we experiment with three sets of prompts: in-domain, out-of-domain, and Spanish. For visualisation specifications we use Vega-Lite's JSON declarative specification [SMWH17] which can then be directly rendered to produce visualisations.

Figure 1 shows an example of a one-shot in-domain prompt for an English user query. Each prompt starts with the instruction to generate a Vega-Lite specification, followed by the completed example. Task metadata is provided to aid the model in producing a relevant output. Each prompt ends with an partially complete solution. The LLM will generate the remaining tokens that are associated with the corresponding Vega-Lite specification.

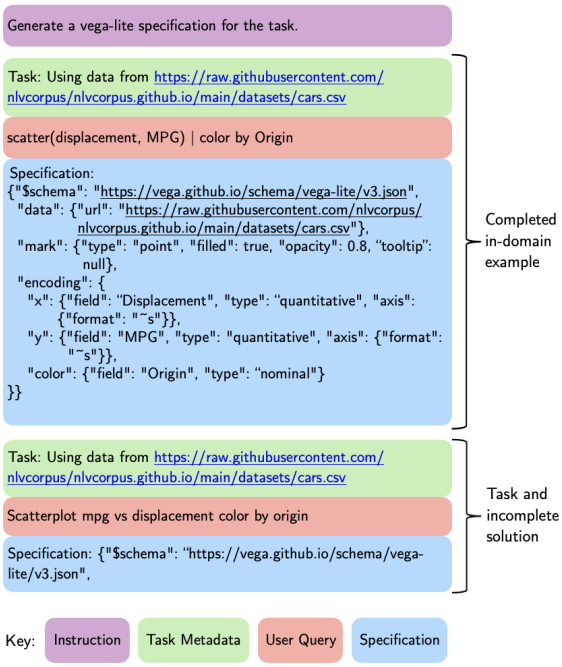

Figure 1: *Example one-shot in-domain prompt for a given English user query.*

3. Initial Results and Future Work

Evaluation protocol. We use a series of unit tests to evaluate the generated specifications. As the output of our models are intended to be valid JSON specifications, we first load the generated specification string into a Python dictionary. If this fails, we categorise these as *Syntax Errors*. If successful, the dictionary is validated through Altair [VGH*18] which compiles the Vega-Lite specifications. If this fails, we categorise these as *Validation Errors*. Finally, the successfully compiled chart is compared to the target chart using Altair. Altair is able to account for differences in ordering of the different chart properties in the specification, allowing for functionally identically charts to be compared. If they do not match, we categorise these as *Logic Errors*. We use *Pass@k* as a metric which represents the proportion of prompts that produced a specification that passed all tests in k samples. Hence, Pass@1 represents the proportion of prompts that successfully generate specification that passes all tests and matches the target visualisation from the corpus.

Results. We present the results in Table 1. We observe that the in-domain prompts tend to produce comparatively better results than the out-of-domain prompts across all metrics. Although over 90% of the out-of-domain prompts result in valid Vega-Lite specification, less than 1% actually matched the target specification. Importantly, we note that the results for Spanish language queries are very similar to the English results despite have no in-context examples in Spanish. We conjecture that our results are likely due to

Prompt Type	Pass@1	Syntax Error	Validation Error	Logic Error
English Out-of-domain	0.12%	4.4%	2.7%	92.8%
English In-domain	81.1%	3.4%	0%	15.5%
Spanish In-domain	78.1%	4.3%	0.12%	17.4%

Table 1: *Pass@1 metric and error breakdown for bloom-560m generated Vega-Lite specifications for different prompt types.*

limitations of the corpus: currently, only 30 distinct visualisations are used, so the in-domain prompt could have an example with the similar target specification as the query. In this case, the model may simply be repeating the tokens that it has already seen. We however note that our results provide a proof of concept that a pre-trained LLM can produce a Vega-Lite specification from a user query in multiple languages using prompt engineering.

Future Work Current datasets are limited, covering only a limited number of visualisation types and are mostly oriented towards English language queries [LTL21; SNL*21]. As a followup, we aim to conduct a larger-scale study to collect natural language that better captures user goals across multiple languages. Current methodologies for evaluating visualisation specifications are *ad hoc* and limited. Generated visualisations are only being compared to one specific goal, which may reject valid visualisations that do satisfy a user's requirements, but differ in some way from the goal. We posit that a human-centred validation of the final visualisations perhaps could be more effective.

References

[BMR*20] BROWN, TOM B., MANN, BENJAMIN, RYDER, NICK, et al. *Language Models are Few-Shot Learners*. arXiv:2005.14165 [cs]. July 2020. DOI: 10.48550/arXiv.2005.14165 1, 2.

[Dib23] DIBIA, VICTOR. *LIDA: A Tool for Automatic Generation of Grammar-Agnostic Visualizations and Infographics using Large Language Models*. arXiv:2303.02927 [cs]. Mar. 2023. DOI: 10.48550/arXiv.2303.02927 1.

[LHJY21] LIU, CAN, HAN, YUN, JIANG, RUIKE, and YUAN, XIAORU. "ADVISor: Automatic Visualization Answer for Natural-Language Question on Tabular Data". *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*. ISSN: 2165-8773. Apr. 2021, 11–20. DOI: 10.1109/PacificVis52677.2021.00010 1.

[LTL*22] LUO, YUYU, TANG, NAN, LI, GUOLIANG, et al. "Natural Language to Visualization by Neural Machine Translation". *IEEE Transactions on Visualization and Computer Graphics* 28.1 (Jan. 2022). Conference Name: IEEE Transactions on Visualization and Computer Graphics, 217–226. ISSN: 1941-0506. DOI: 10.1109/TVCG.2021.3114848 1.

[LTL21] LUO, YUYU, TANG, JIAWEI, and LI, GUOLIANG. *nvBench: A Large-Scale Synthesized Dataset for Cross-Domain Natural Language to Visualization Task*. arXiv:2112.12926 [cs]. Dec. 2021. DOI: 10.48550/arXiv.2112.12926 2.

[MS23] MADDIGAN, PAULA and SUSNJAK, TEO. *Chat2VIS: Generating Data Visualisations via Natural Language using ChatGPT, Codex and GPT-3 Large Language Models*. arXiv:2302.02094 [cs]. Feb. 2023. DOI: 10.48550/arXiv.2302.02094 1.

[NSS21] NARECHANIA, ARPIT, SRINIVASAN, ARJUN, and STASKO, JOHN. "NL4DV: A Toolkit for Generating Analytic Specifications for Data Visualization from Natural Language Queries". *IEEE Transactions on Visualization and Computer Graphics* 27.2 (Feb. 2021). Conference Name: IEEE Transactions on Visualization and Computer Graphics, 369–379. ISSN: 1941-0506. DOI: 10.1109/TVCG.2020.3030378 1.

[SMWH17] SATYANARAYAN, ARVIND, MORITZ, DOMINIK, WONGSUPHASAWAT, KANIT, and HEER, JEFFREY. "Vega-Lite: A Grammar of Interactive Graphics". *IEEE transactions on visualization and computer graphics* 23.1 (2017). Publisher: IEEE, 341–350 2.

[SNL*21] SRINIVASAN, ARJUN, NYAPATHY, NIKHILA, LEE, BONGSHIN, et al. "Collecting and Characterizing Natural Language Utterances for Specifying Data Visualizations". *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. New York, NY, USA: Association for Computing Machinery, May 2021, 1–10. ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445400 1, 2.

[SS17] SRINIVASAN, ARJUN and STASKO, JOHN. "Natural language interfaces for data analysis with visualization: considering what has and could be asked". *Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Short Papers*. EuroVis '17. Goslar, DEU: Eurographics Association, June 2017, 55–59. DOI: 10.2312/eurovisshort.20171133 1.

[TS19] TORY, MELANIE and SETLUR, VIDYA. "Do What I Mean, Not What I Say! Design Considerations for Supporting Intent and Context in Analytical Conversation". *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. Oct. 2019, 93–103. DOI: 10.1109/VAST47406.2019.8986918 1.

[TT20] TIEDEMANN, JÖRG and THOTTINGAL, SANTHOSH. "OPUS-MT – Building open translation services for the World". *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisboa, Portugal: European Association for Machine Translation, Nov. 2020, 479–480 2.

[VGH*18] VANDERPLAS, JACOB, GRANGER, BRIAN, HEER, JEFFREY, et al. "Altair: Interactive Statistical Visualizations for Python". *Journal of Open Source Software* 3.32 (2018). Publisher: The Open Journal, 1057. DOI: 10.21105/joss.01057 2.

[VSP*17] VASWANI, ASHISH, SHAZEER, NOAM, PARMAR, NIKI, et al. *Attention Is All You Need*. arXiv:1706.03762 [cs]. Dec. 2017. DOI: 10.48550/arXiv.1706.03762 1.

[WSF*23] WORKSHOP, BIGSCIENCE, SCAO, TEVEN LE, FAN, ANGELA, et al. *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. arXiv:2211.05100 [cs]. Mar. 2023. DOI: 10.48550/arXiv.2211.05100 1.

[WTB*22] WEI, JASON, TAY, YI, BOMMASANI, RISHI, et al. "Emergent abilities of large language models". *Transaction of Machine Learning Research* (2022) 1.

[WWS*21] WU, AOYU, WANG, YUN, SHU, XINHUAN, et al. *AI4VIS: Survey on Artificial Intelligence Approaches for Data Visualization*. arXiv:2102.01330 [cs]. July 2021. DOI: 10.48550/arXiv.2102.01330 1.