The Lay of the Land: Data Visualizations of the Language Data and Domains of Wikidata *

Niel Chah[1][0000−0002−3377−7823] and Periklis Andritsos[2]

University of Toronto, Faculty of Information
[1]niel.chah@mail.utoronto.ca,[2]periklis.andritsos@utoronto.ca

Abstract. Wikidata is a well-known collaborative knowledge graph containing multilingual data for hundreds of language locales, from Arabic (ar) to Zulu (zu). However, research on the state of language localization and the full extent of the topics that currently exist on Wikidata has been relatively limited. This paper presents data visualization in the form of various heatmaps resulting from machine learning experiments using BERT embeddings, t-SNE, and K-means clustering that show the kinds of topics, or domains, in Wikidata and the frequency of multilingual data across such domains. Such heatmaps are useful for identifying domains where language data is lacking and should be populated. A representative heatmap is also presented that shows the localization across domains for the top 30 languages geographically from west to east. Finally, future work in this space is also considered.

Keywords: Wikidata · Multilingual data · Domains · Machine learning · Data visualization

This is a poster submission for ISWC 2021.[1]

1 Introduction

Wikidata is a "free and open knowledge base that can be read and edited by both humans and machines" that is actively researched and supports hundreds of languages [3]. For instance, the Wikidata *item* for the "Moon" is "Q405"[2], and is linked to labels, descriptions, and aliases in various languages. As KGs like Wikidata are increasingly used in new products and AI services, it is worth further looking into Wikidata's data coverage across international languages and communities. At a high level, this paper presents developing work that seeks to answer the following research question in order to offer a novel look into the multilingual data on Wikidata. What is the distribution of multilingual data across different *languages* and *topics or domains* in Wikidata, and how can such distributions identify lack of coverage in multilingual data?

* Supported by funding from NSERC.

[2] https://www.wikidata.org/wiki/Q405

2 Related Works

To our knowledge, existing tools have not organized Wikidata into machine learned subject matter topics, or domains, and visualized their distribution of multilingual data. Other related tools and dashboards describe the multilingual information on Wikidata, with a list of such tools available on Wikidata.[3] For instance, WDProp, is a dashboard that visualizes the frequency of translated label, description, and alias values in the Wikidata ontology, but not other items.[4]

3 Methodology

The March 1, 2021 Wikidata JSON dump (>80GB) was profiled using Python.

Generating P31 object sets. The P31 property ("instance of") is a useful way to determine the *identity* of Wikidata items as it accounts for 8.07% of all statements in Wikidata and 96.48% of items have at least one P31 statement. P31 object sets, where $s \in S$, were generated. Based on triples in a $(Qid, P31, object)$ format, where the item's "Qid" is the subject and the object is also another item, a P31 object set is a set of the Qid(s) in the object positions for all P31 triples (e.g. `[(Q13442814) scholarly article, (Q7318358) review article]`). As the distribution of sets has a long tail, a threshold of at least one thousand items was used which resulted in a $|S|$ of 1588 sets that account for 97.15% of all items.

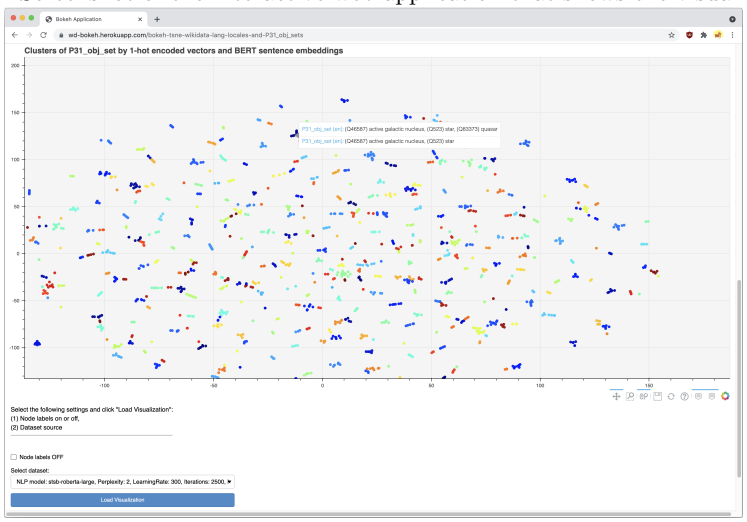
For one experiment, the composition of the P31 object sets were converted into 1-hot encoded vectors as follows. For each set, s_n, which can be composed of multiple Wikidata items $q_1, q_2, ..., q_n$, a feature vector, $\mathbf{f_q}$ is created of length $|S|$ where the value of $\mathbf{f}_{\mathbf{q}_i}$ is one if the set contains the Wikidata item q_n and zero otherwise. In another approach, the English labels and descriptions were collected for each of the items q_n in a set s_n and put into a vector for natural language processing (NLP). These NLP features were encoded into 768-dimensional sentence embeddings by Sentence-BERT models [2]. After experimentation, the `stsb-roberta-large` model was chosen due to its optimization for Semantic Textual Similarity (STS) comparison.

Hyperparameter tuning of t-SNE and K-means. The t-Distributed Stochastic Neighbor Embedding (t-SNE) technique was used to reduce the data to two dimensions, followed by K-means clustering. In order to achieve optimal clustering, hyperparameter tuning of t-SNE *perplexity*, *learning rate*, and *number of iterations* and K-means *number of clusters* using grid search was used to obtain the highest silhouette scores. Using 1-hot encoded vectors, perplexity of 2, learning rate of 150, number of iterations of 15,000, and $k = 65$ clusters produced the highest silhouette score of 0.868. Using BERT embeddings, the optimal parameters of perplexity of 2, learning rate of 300, number of iterations of 2500, and $k = 375$ clusters resulted in silhouette score of 0.714.

[3] `https://www.wikidata.org/wiki/Wikidata:Tools/Visualize_data`
[4] `https://wdprop.toolforge.org/wdprop.html`

An interactive web application tool was created to visualize the results of the optimal hyperparameters and other experimental attempts using the interactive Python library bokeh, GitHub, and Heroku.[5]

Fig. 1. Screenshot of the interactive web application that shows the visualization.



4 Findings

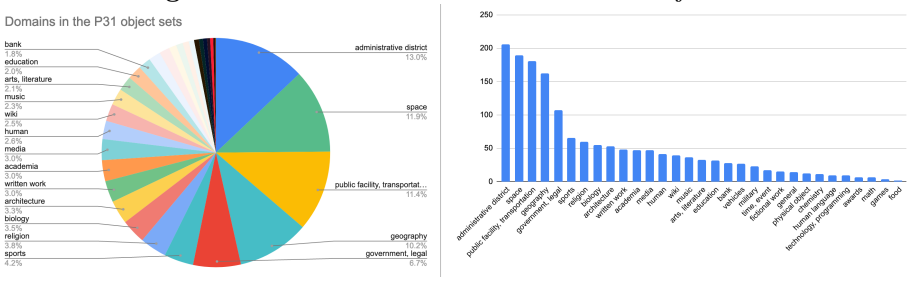
Domains of Wikidata. Wikidata's ontology does not explicitly assign classes of entities into specific subject matter "domains" as was the case for Freebase with such domains as "awards" and "zoos" [1]. For external validation of the learned clusters, ground truth labels were created by iteratively generating category labels by the authors using a grounded theory approach to qualitatively code each P31 object set based on its English labels, descriptions, aliases and other relevant information. The initial seed of subject matters was inspired by examining word clouds of the clusters and finding overarching themes. To address the subjective nature of these assignments, new domains were created for generally different and orthogonal subjects.

The distribution of the different domains in the ground truth labeled data is shown in Figure 2. Many of the domains found in the P31 object sets concern administrative districts (e.g. `(Q18524218) canton of France`), space or astronomy (e.g. `(Q2488) spiral galaxy`), public facilities or transportation (e.g. `(Q28564) public library`, `(Q2175765) tram stop`), and geography (e.g. `(Q4022) river`).

To check the external validity of the clusters with the ground truth labels $t_j \in T$, the *purity* measure was used. $Purity = \frac{1}{N}\sum_{i=1}^{k} max_j |c_i \cap t_j|$, is calculated

[5] `https://wd-bokeh.herokuapp.com/`

Fig. 2. Distribution of the domains in the P31 object sets.



by assigning each cluster to the class which is most frequent in the cluster c_i and measuring the accuracy of this assignment by counting the number of correct assignments and dividing by N. A purity metric of 0.799 was achieved for the BERT embedding clusters, compared to 0.352 for 1-hot encoded vectors. These metrics indicate that the BERT-based sentence embeddings were able to learn more internally consistent clusters.

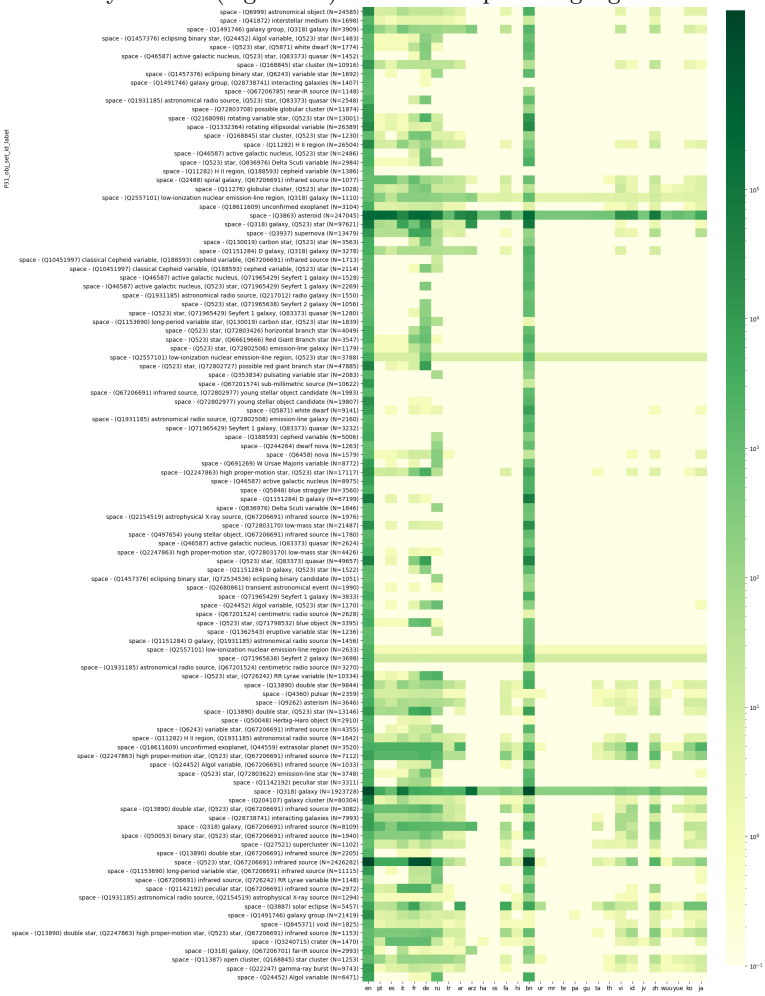
Heatmaps of multilingual data across domains. Heatmaps were generated using the Python library seaborn to depict the prevalence of localized language data across the various domains of Wikidata. The heatmaps may be adjusted to show counts (best viewed on a log-scale due to large value ranges), and percents of total localized in order to identify areas where Wikidata labels, aliases, and descriptions are lacking and open to focused data augmentation and localization efforts. These various heatmaps are accessible at the aforementioned web application tool.[6] A representative example of these heatmaps is shown in Figure 3, where the top 30 languages by the number of global speakers according to *Ethnologue*[7] are arranged from left to right to correspond with west to east geographically.

5 Future Work

Future work that builds on these findings will utilize the generated heatmaps to determine areas of missing multilingual data in Wikidata. Instead of general statistics on the prevalence of the most or least populated languages, by using the heatmaps specific parts of Wikidata can be identified for targeted data augmentation and enrichment of labels, aliases, and descriptions. For instance, knowledge graph embeddings and link prediction may be directed at the most interesting portions of Wikidata that may be identified.

[6] `https://wd-bokeh.herokuapp.com/`
[7] `https://www.ethnologue.com/guides/ethnologue200`

Fig. 3. A portion of the overall heatmap of localized multilingual data for labels in the "space" astronomy domain (log-scaled) across the top 30 languages.



References

1. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data (2008)
2. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on EMNLP (11 2019), `http://arxiv.org/abs/1908.10084`
3. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (Sep 2014). https://doi.org/10.1145/2629489, `http://dl.acm.org/citation.cfm?doid=2661061.2629489`