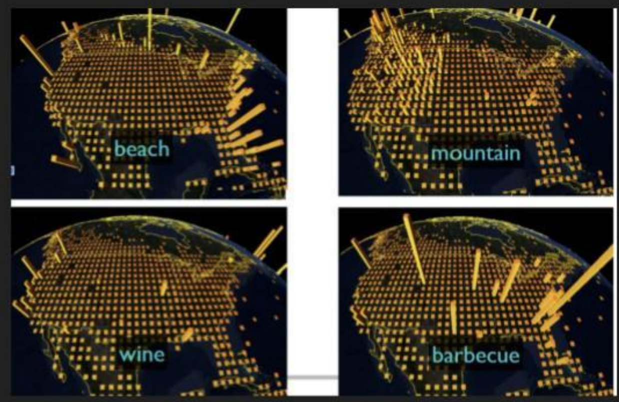
Transformer Based Geocoding

Yuval & Vitaly NLP group Rafael Ltd

USA Tweets words distribution

Looks like Some words have distinct spatial distribution.

Can we use text data with location labels to predict location from free text ?



Geocoding model recipe

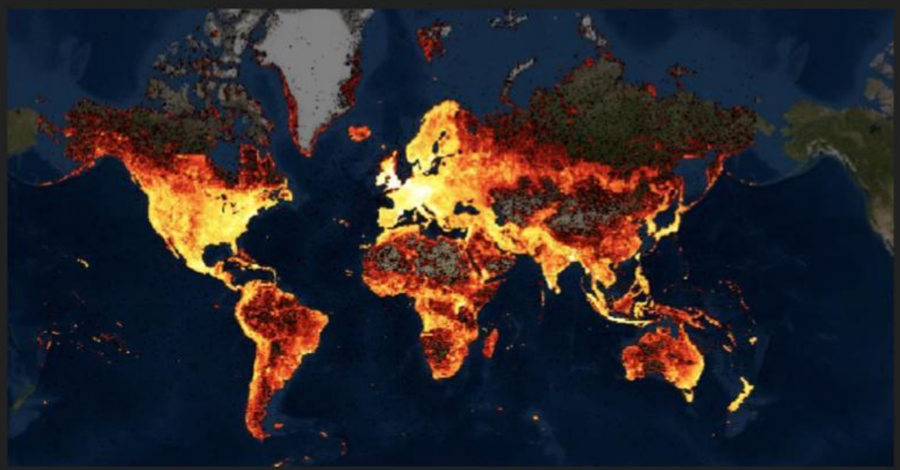
- Data: free text with location labels

Id	Latitude	Longitude	Text
01	53.96	-1.08	historic county of England
02	51.0	10.0	country in Central Europe
03	35.88	14.5	sovereign state in Southern Europe
04	57.30	-6.36	whisky distillery in Highland, Scotland, UK
05	47.39	0.69	city and commune in Indre-et-Loire, Centre-Val de Loire, France
06	-33.0	-71.0	sovereign state in South America

Wikidata location - 8M records

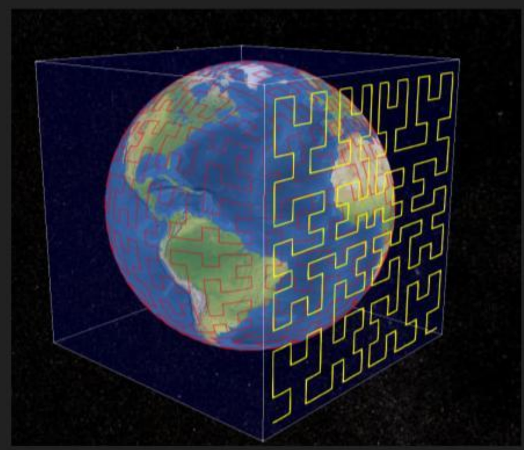
- Downstream task: Area classification? Categories? Grid? Resolution?

Wikidata spatial distribution



S2 Projection data labeling

First digit: the cell cube face with a digit between 0 to 5



Location data labeling

Next digits represent the corresponding node in the quad tree with a digit between 0 to 3

Table 1: S2 Geometry levels.

Level	Average area	Number of cells
00	85M km^2	6
01	21M km^2	24
02	5M km^2	96
03	1.3M km^2	384
04	330K km^2	1536
05	83K km^2	6K
06	20K km^2	24K
07	5K km^2	98K
08	1297 km^2	393K
09	324 km^2	1573
10	81 km^2	6M
..	..	..
29	2.95 cm^2	$1729 * 10^{15}$
30	0.74 cm^2	$7 * 10^{18}$

Adaptive Cell Partitioning

Split the earth along the queued tree until all cells contains less than X* samples

- balanced class distribution
- Parameter efficiency - model capacity is spent on densely populated areas

* X == max samples per cell (hyper parameter)

S2 Geometry adaptive partitioning of our dataset.

Location data label encoding Examples

First digit: the cell cube face with a digit between 0 to 5

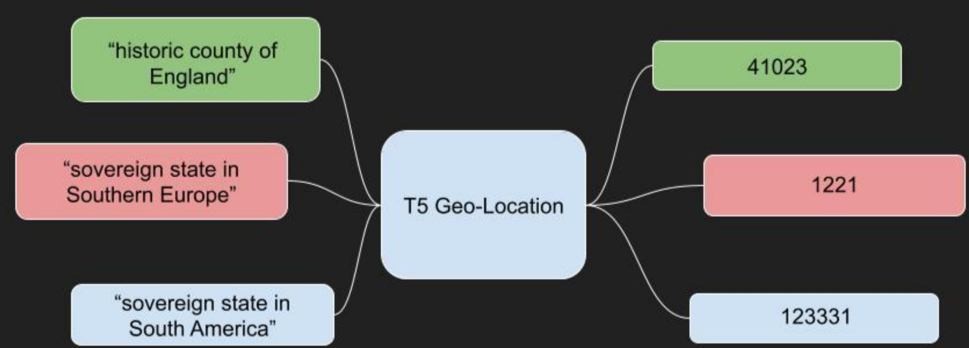
Next digits represent the corresponding node in the quad tree with a digit between 0 to 3

Cell description	Cell representation
Face cell 2	2
Subcell 2 of face cell 1	12
Subcell 1 of subcell 3 of face 4	431

lake in Eksjö Municipality, Sweden	20302303
ancient monument in Denmark (2976)	20331122

Geocoding Sequence-to-Sequence Model

- Based on the T5-base transformer architecture
- Text records as input and the location cell encoding as output



A diagram of our text-to-location framework

Results

- Inference examples - true and predicted labels:

Text	Predicted Label	True Label
townland in Drummaan, County Clare, Ireland	21002321	21002321
lake in Eksjö Municipality, Sweden	20302303	20302303
ancient monument in Denmark (2976)	20331122	20331122
school in Cheshire West and Chester, UK	210033112	210033113
mountain in Iran	1333313	133302
railway stop in Harburg, Germany	20331203	20331022

- Evaluation results:

Evaluation metric	Results
Flat accuracy	0.51547
Hierarchy accuracy	0.791

Evaluation Metric

- Which metric to choose? Accuracy measure? Mean distance error?
  - Both fail to capture the inherent hierarchical nature of the label.
- Hierarchical classification metric
  - hierarchical precision (hP): $$hP = \frac{\sum_i |P_i \cap T_i|}{\sum_i |P_i|},$$
  - hierarchical recall (hR): $$hR = \frac{\sum_i |P_i \cap T_i|}{\sum_i |T_i|},$$
  - and hierarchical f-measure (hF): $$hF = \frac{2 * hP * hR}{hP + hR}$$

* P_i is the set consisting of the most specific class predicted for each test example i, and all of its ancestor classes. T_i is the set consisting of the true most specific class of test example i, and all its ancestor classes. Each summation is computed over all of the test set examples.

Transformer Based Geocoding

Yuval Solaz* yuval.solaz@gmail.com
Vitaly Shalumov* vitaly.shalumov@gmail.com

January 4, 2023

Abstract

In this paper, we formulate the problem of predicting a geolocation from free text as a sequence-to-sequence problem. Using this formulation, we obtain a geocoding model by training a T5 encoder-decoder transformer model using free text as an input and geolocation as an output. The geocoding model was trained on geo-tagged wikidump data with adaptive cell partitioning for the geolocation representation. All of the code including Rest-based application, dataset and model checkpoints used in this work are publicly available.

1 Introduction

Social media such as Twitter and Wikipedia contains considerable amount of location-related text data. In this paper, we develop a model that learns to predict spatial probabilities from free text. Given a query sentence, the model outputs a discrete probability distribution over the surface earth, by assigning each geographical cell a likelihood that the input text relates to the location inside said cell. The resulting model is capable of localizing a large variety of sentences. Viewing the task as a hierarchical classification problem allows the model to express its uncertainty in the location associated with the text. The resulting model can be used for resolving ambiguity of the location references in the text. This capability is central to the success of finding exact location from free text. For example, *Paris* can refer to more than one possible location. In a context such as: *The International Olympic Committee confirmed the city chosen to host the Olympic Games in 2024. The Games will be held in Paris*, geocoding models like the one proposed in this paper can help in the resolution of the correct location.

This work introduces the following contributions:

- Synthesizing a dataset for supervised learning, including adaptive cell partitioning.
- Formulating the geocoding problem as a sequence-to-sequence problem.
- Training an end-to-end geocoding model using said formulation.
- Publicly releasing the curated dataset, a Rest-based application and the T5 geocoding model.

*Equal Contribution