GEOSPATIAL DATA UNDERSTANDING:

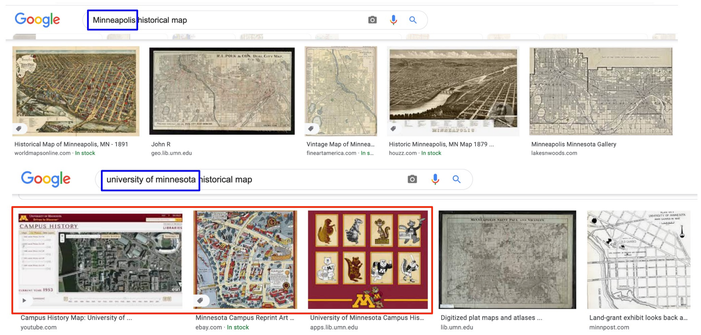
A Peek into Historical Maps and Contemporary Geospatial Databases

Zekun Li, Department of Computer Science and Engineering

University of Minnesota
Driven to Discover®

Introduction

- Historical maps offer a wealth of valuable information of our past, **millions** of scanned maps are made widely **available** nowadays.
- But most of the maps remain **unanalyzed**
- **Reason:** map processing is **time-consuming** and **costly**



Search result of "Minneapolis historical map" is reasonable, but result for "**University of Minnesota**" is **erroneous**

We want to develop a machine-learning method to **read the historical map** and **establish connections to** contemporary geospatial databases!

Challenges

- Historical maps looks quite **different from natural scene images**, spotting models trained on general domain data does not perform well on maps

Sample images from ICDAR 15 dataset

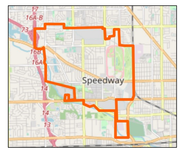
Sample image patches from David Rumsey historical map collection

- Linking to contemporary geospatial database can be challenging due to the usage of **same places names** in different locations

Geo-entity Feature Representation

- Most geo-entities exist as **point** data (e.g. GeoNames).
- Geo-entity names can be **ambiguous** without knowing the geometry

SpaBERT Repo



A **town** in Marion County, Indiana

A **road** in Los Angeles, California

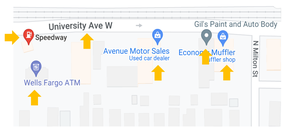
A **gas station** in Minneapolis, Minnesota

Q: What is **Speedway**?

We shall know the **characteristics** of a geo-entity by its **surrounding entities**, similar to knowing word meanings by their linguistic context.

Pacific Avenue
30th Place
Speedway

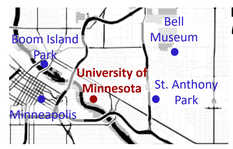
Speedway should be a **road**!



[*Car-related*]
Avenue Motor Sales
Gil's Paint and Auto Body
[*Common near gas stations*]
Wells Fargo ATM

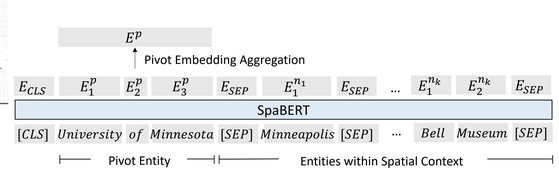
Speedway should be a **gas station**!

- Problem Setting & Approach
  - Input: Geo-entity **name** and **point location** (image coord. or geo-coord.)
  - Goal: Produce **general-purpose** geo-entity feature representation



Pivot entity
Entities within spatial context

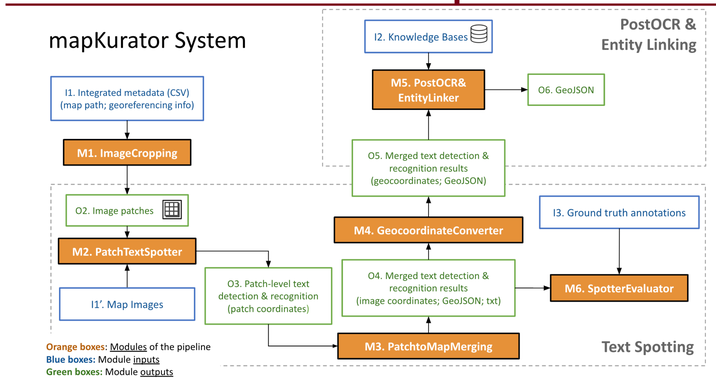
Neighbor geo-entities sort by distance: $(dist \uparrow)$
University of Minnesota, Minneapolis, St. Anthony Park, Bloom Island Park, Bell Museum



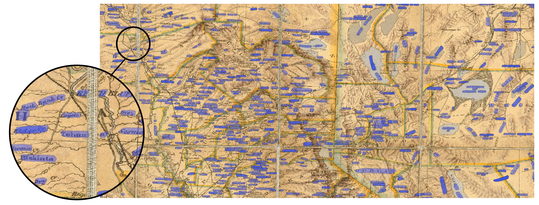
- Downstream task: **Geo-entity Linking**
  - **Task**: Link geo-entities in scanned historical maps (USGS) to Wikidata
  - ***Setting***: USGS map entities are associated with **pixel** coordinates; Wikidata entities are associated with **geo-coordinates**

Model	MRR	R@1	R@5	R@10
BERT$_{Base}$	.400	.289	.559	.635
RoBERTa$_{Base}$	.326	.232	.446	.540
SpanBERT$_{Base}$	.164	.138	.201	.213
LUKE$_{Base}$	.306	.188	.440	.547
SimCSE$_{BERT\text{-}Base}$	.453	.371	.547	.628
SimCSE$_{RoBERTa\text{-}Base}$	.227	.188	.264	.301
SpaBERT$_{Base}$	.515	.338	.744	.850

Automatic Historical Map Understanding

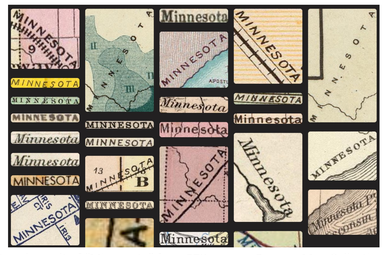


- **Inputs:** Historical map images (.png/.geotiff) **or** metadata providing map path
- **Outputs:** Recognized text labels **&** label bounding polygons **&** Identifier to OSM

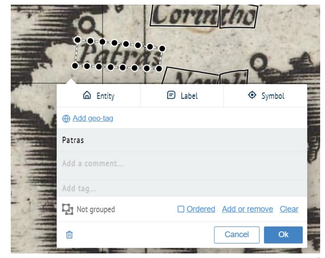


mapKurator Website

Recognized text labels from "*Map Of California And Nevada*" by *Geological Survey of California*



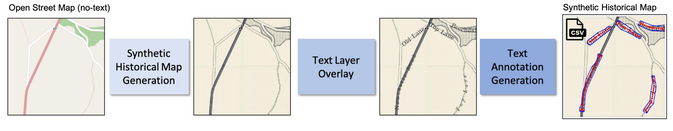
Search result of "Minnesota" from 57K maps in Rumsey Map Collection



Display mapKurator spotting result in Recogito web interface

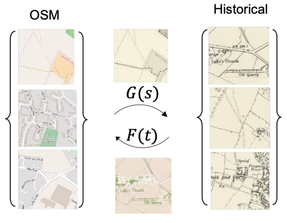
Generate Synthetic Historical Maps

- Gathering **training data** for historical maps is important, while manual annotation takes a lot of time and effort
- We propose to generate **synthetic historical maps** to aid the training of text detection models
- **General Idea:**
  - Create synthetic map **background without** any **text labels**
  - Automatically **place text labels** and compute ground-truth annotation (of text bounding polygon)



SynthMap Demo

- **Source** map: Clean (no text) OpenStreetMap tiles to provide background
- **Target** map style: **Ordnance Survey** 6-inch map during year 1888-1913
- **Model**: CycleGAN to efficiently perform **style transfer**



Adversarial Loss:

$$\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{\text{data}}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(1 - D_Y(G(x))]$$

Cycle Consistency Loss:

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1].$$

*Details for automatically computing the annotation info can be found in the paper

Conclusion

- **SynthMap**, a dataset of synthetic historical map images generated from OSM tiles using cycleGAN to help improve text detection.
- **SpaBERT**, a BERT-based language model to capture the relations between 2D geo-entities and produce spatial-context-aware features.
- **mapKurator**, a machine learning system for historical map understanding.

References

[1] Li, Zekun. "Generating historical maps from online maps." *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 2019.
[2] Li, Zekun, et al. "Synthetic map generation to provide unlimited training data for historical map text detection." *Proceedings of the 4th ACM SIGSPATIAL GeoAI Workshop*. 2021.
[3] Li, Zekun, et al. "SpaBERT: A Pretrained Language Model from Geographic Data for Geo-Entity Representation." *Proceedings of the EMNLP*. 2022.
[4] Li, Zekun, et al. "An automatic approach for generating rich, linked geo-metadata from historical map images." *Proceedings of the 26th ACM SIGKDD*. 2020.

Acknowledgement

National Endowment for the Humanities
David Rumsey Map Collection — Cartography Associates — Powered by Luna Imaging
The Alan Turing Institute
Recogito | partner of the Pelagios network
Knowledge Computing Lab — University of Minnesota