Automatic Extraction of Phrase-Level Map Labels from Historical Maps

Haowen Lin (Mentor: Yaoyi-Chiang)

Spatial Sciences Institute, University of Southern California

Problem Statement

Textual information from historical maps consists of useful information

The result of OCR scanning or manual map digitization are separated word bounding boxes

This limits the opportunity for:

- prevents historical maps from being indexed and searched by meaningful phrases (e.g., Sand hills is more useful than hills)
- prevents researchers to form useful gazetteer

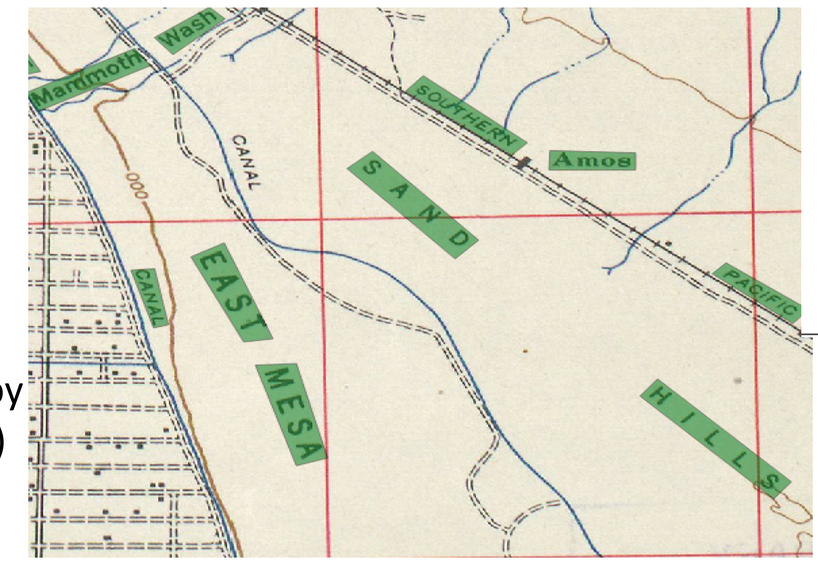
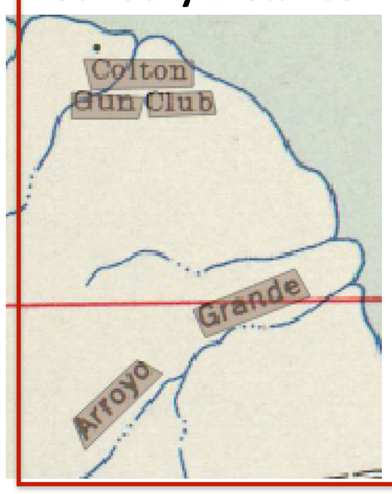


Table 1: Input Data and Output Data for Polygons in Figure 1

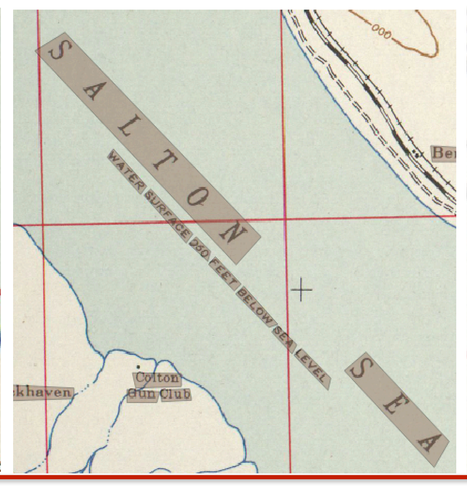
Input Data (Geo polygon)	Output Data
Mammoth	Linking with "Wash"
Wash	Linking with "Mammoth"
EAST	Linking with "MESA"
MESA	Linking with "EAST"
SAND	Linking with "HILLS"
HILLS	Linking with "SAND"
SOUTHERN	Linking with "PACIFIC"
Amos	No linkage
PACIFIC	Linking with "SOUTHERN"

Linking Word Approach: Feature Generation

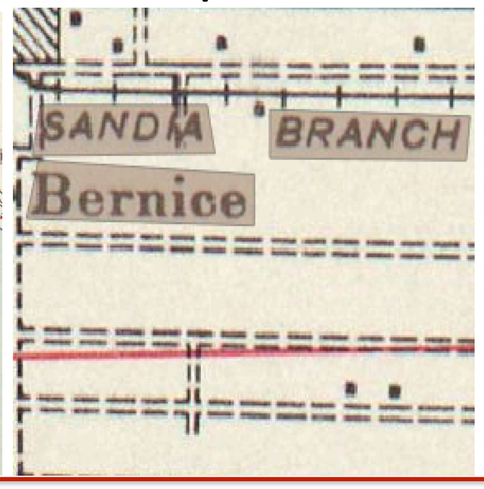
Boundary Distance



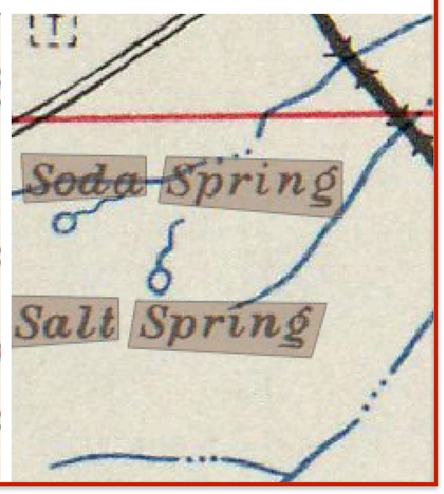
Text Area for Each Character



Capitalization



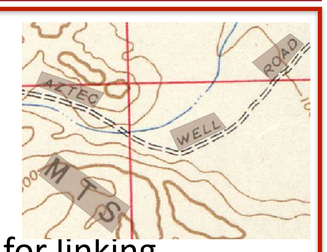
Textual Content



Feature Generation Challenges

Curved Polygons

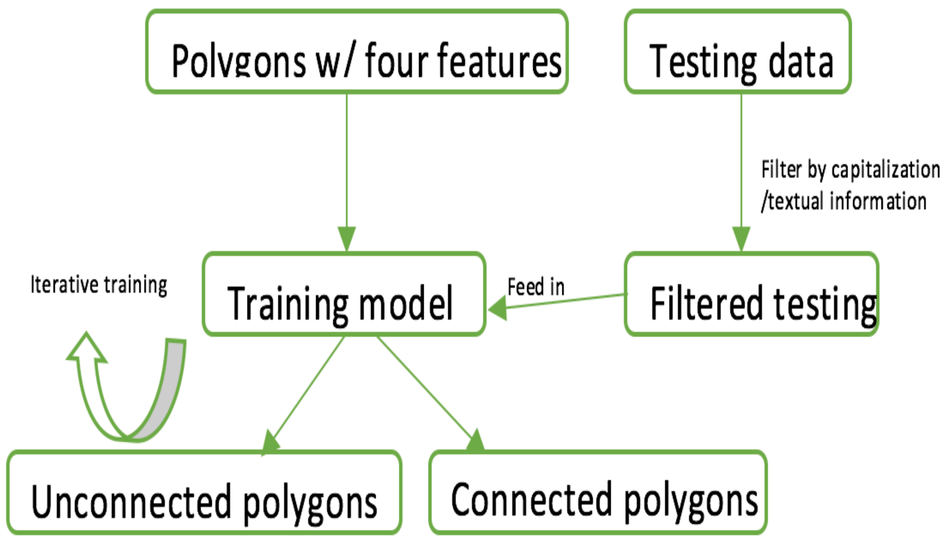
- Orientation matters
- Minimum bounding box could decide orientation



Textual content for linking

Support Vector Machine

Finding the "correct" threshold for each of the features in to connect single words would not be reliable, and this problem has an intuitive implication to use Support Vector Machines (SVMs) in the classification settings.



Data Flow Chart

Preliminary Result

Table 2: Experiment results

Result	USGS 60 Inches-Salton	Ordnance Survey 60 Inches	USGS 15 Inches-Brawley
Precision	91.56%	91.67%	79.31%
Recall	40.42%	38.60%	32.85%
Total Phrases	95	84	134

Challenges:

- Unbalanced data influence precision and recall
- Polygons with huge distance/ text area differences are still connected

Conclusions

- present an algorithm that combines textual and spatial information of map words to automatically generate meaningful place information
- Future work include generating more features for evaluation and trying other types of machine learning algorithms to deal with the situation when the map label is curved

References:
[1] YY. Chiang. Unlocking Textual Content from Historical Maps- Potentials and Applications, Trends, and Outlooks. International Conference on Recent Trends in Image Processing and Pattern Recognition, pages 111-124. Springer, 2016.
[2] M. Heliński, K. Miłosz, and P. Tomasz. Report on the comparison of Tesseract and ABBYY FineReader OCR engines. 2012.