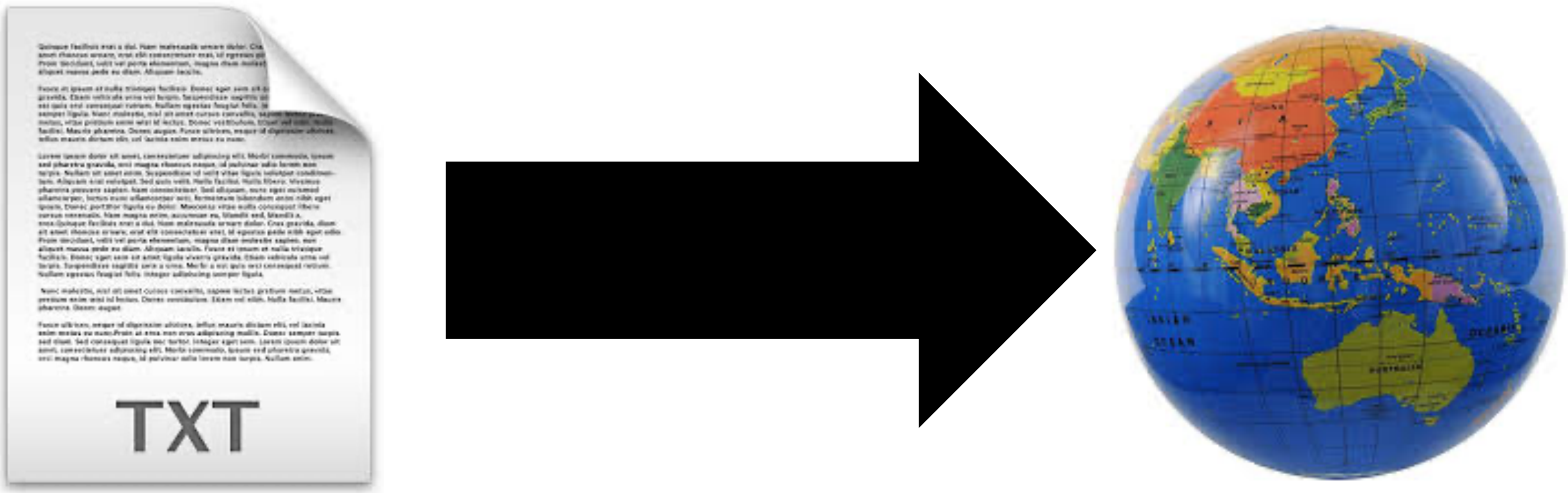
Challenges in resolving place names over text

Bruno Martins

University of Lisbon, Instituto Superior Técnico and INESC-ID

Place reference resolution

- Two separate sub-problems:
  - Recognizing place references in textual documents
  - **Resolving the recognized place references into unambiguous locations**
- Related task: *Document geolocation (not covered in my talk)*



Many possible applications

(e.g., in the humanities and the social sciences)

- **Search** within document collections according to geospatial constraints
    - Geographic information retrieval, question answering, etc.

- **Visualization** of topics discussed within textual documents over maps
    - Explore textual information with techniques from thematic cartography

- **Spatial analysis** leveraging information originally encoded in text
    - Geographical text analysis based on collocations with places
    - Methods from traditional spatial analysis (e.g., hotspot detection, clustering, etc.)

Challenges in language understanding

- **Ambiguity in matching geographic names (geo/geo)**
  - Different places sharing the same name
    - *Dallas* in Texas versus *Dallas* County in Alabama

- Ambiguity between geographic and non-geographic names (geo/non-geo)
  - Places names that frequenty also have other non-geographic meanings
    - Person named *Charlotte* versus *Charlotte* County in Virginia
  - Not going to be addressed in this talk, given that this concerns place reference recognition

- **Reference ambiguity**
  - Places can be refered to through different names
    - Names like *Big Apple* or *New Amsterdam* can both be used to refer to *New York City*

- Many other challenges not addressed in this talk
  - References to approximate/vague locations (e.g., that use distance and/or direction qualifiers)

One approach to this problem

- Leverage existing **named entity recognition** tools to perform place reference recognition in the textual sources
  - Pretrained Transformer encoder models (e.g., BERT) plus improvements (e.g., CRF layer)

- **Neural models** for representing recognized place references, plus surrounding context, and **directly infering geospatial coordinates** from the representations
  - More traditional approach would involve matching text representations against entries in a gazetteer (i.e., a database associating place names to geospatial coordinates)

- **Possible extensions**, such as matching the textual context representations against external information (e.g., from Earth Observation products)

Overview

- Introduction
- **Previous approaches**
- A neural method for toponym resolution
- Results
- Extensions to base model
- Conclusions and future work

Heuristic methods

- Depend on gazetteers describing locations
  - GeoNames, Wikidata, DBPedia, … and also specialized collections
  - Associate place names to geospatial coordinates (plus other relevant attributes)
- Match place references in text against gazetteer entries (i.e., the candidates for disambiguation)
  - String similarity metrics can be used in this context
- **Use heuristics to decide which entry is the most likely match:**
  - Highest population density
  - Accoring to frequency by which the reference matches the candidate place
  - Promote spatial minimality
  - Promote proximity towards non-ambiguous place references given in surrounding context
  - Many other possibilities (e.g., one sense per discourse)
- Combination of heuristics manually defined through expert knowledge (and/or **trial and error**)

Feature-based supervised learning

- Model the task as a ***learning to rank*** problem
- Similar to more traditional *general-domain* entity linking systems

- Retrieve initial set of candidates from a gazetteer
- Represent each candidate through a set of descriptive features
  - Similar to the scores provided by the heuristic methods
    - Features intrinsic to the candidate (e.g., population density)
    - Features measuring association between reference(+context) and the candidate
- Use a learned model to **compute a *matching score***
  - Models based on ensembles of decision trees are popular
  - Choice of loss function is usually an important aspect to consider
- **Rank candidates according to matching score** (*and optionally decide if NIL*)

Geodesic grids and language models

- Approaches that **bypass the need for a gazetteer**, but that usually require **larger amounts of annotated training data**
  - Text associated to geographic locations (e.g., from Wikipedia pages)
  - Can better deal with incomplete gazetteers, approximate references, etc.

- Partition the study region into multiple (usually small) sub-divisions
  - Use a regular geodesic grid
  - Other partitionings are possible (e.g., quadtrees, overlapping areas, etc.)

- Build a language model for each region, with basis on the available training data
  - E.g,, n-gram language models can be used

- Evaluate the text from the place reference (plus the context) with the language models
- **Pick the region whose language model is more likely to generate the text for the candidate**
  - Optional: perform interpolation from regions that are most likely

Neural network approaches

- Model the problem as a *learning to rank* task:
  - Match neural representations for reference+context and for candidates
  - Similar to modern general entity linking systems
  - Can use pre-trained Transformer encoder models (e.g., BERT)

- **Approaches that directly attempt to predict geospatial coordinates:**
  - Avoid the need for gazetteers (advantages of *language modeling* methods)
  - We will discuss one method like this next!

Overview

- Introduction
- Previous approaches
- **A neural method for toponym resolution**
- Results
- Extensions to base model
- Conclusions and future work

A neural toponym resolution method (1)

- Leverage state-of-the-art approaches for representing text
  - Recurrent neural networks for summarizing sequences of text
  - Pretrained contextual word embedding models (e.g., ELMo or BERT)

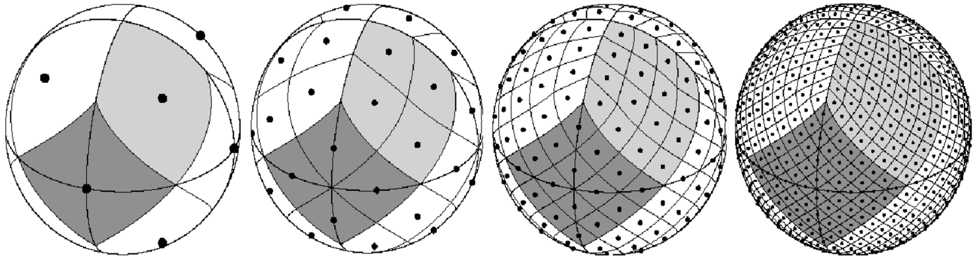
- Model the task as a supervised classification problem
  - Assign place reference (plus the context) to a geospatial region
  - Predict geospatial coordinates with basis on centroids for the regions
  - Use geospatial proximity to complement the standard classification loss

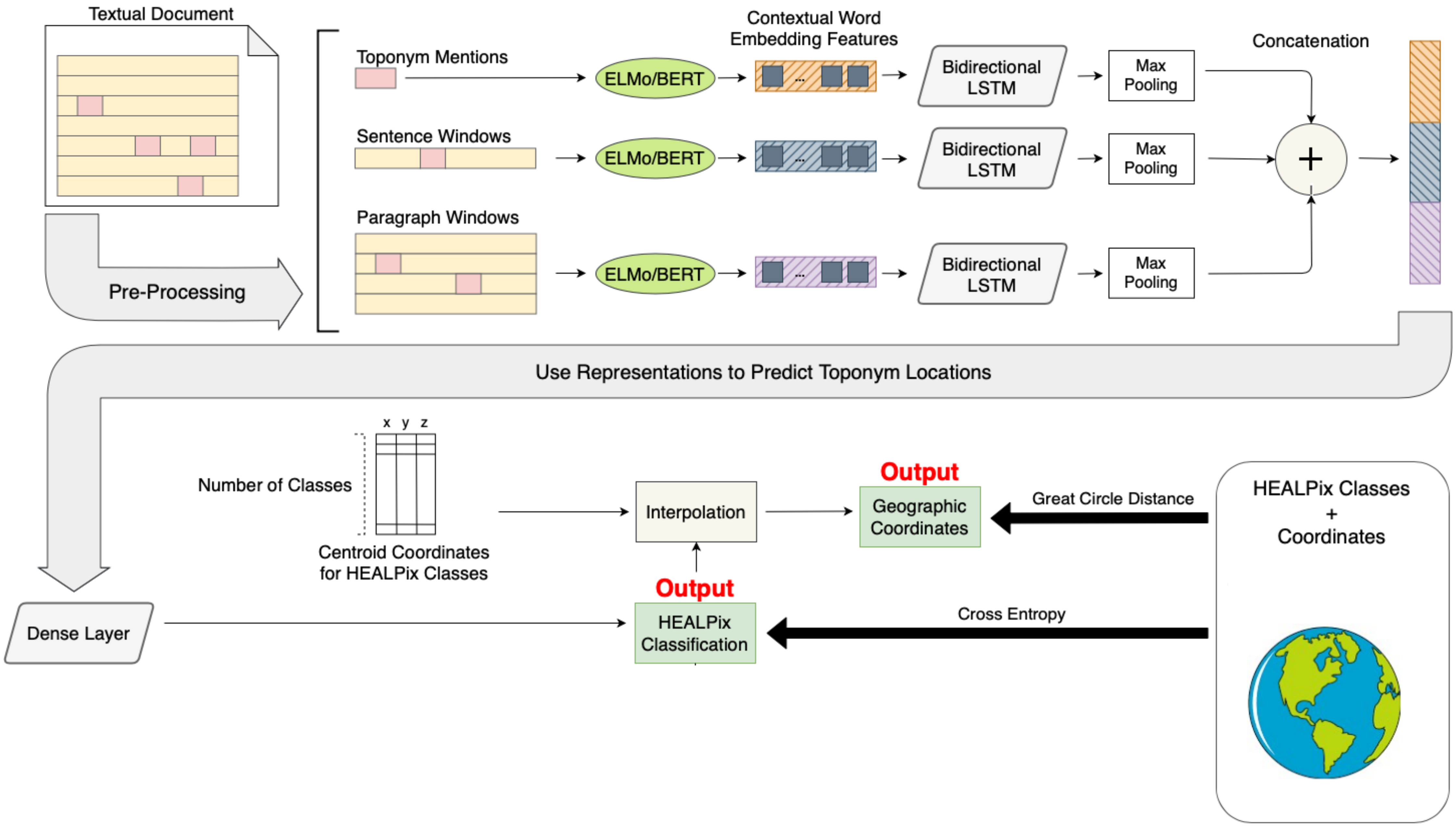
A neural toponym resolution method (2)

- Issues related to building representations for text
    - One can fine-tun ELMo/BERT models (instead of using as feature extractors)
    - Method based on LSTMs is computationally less expensive
    - Context windows are a possible parameter to tune

- Issues related to the geospatial partitioning of the study region
    - Can use alternatives to a regular grid (e.g., quadtree partitioning)
    - HEALPix regions provide interesting properties for large study areas

- Many possible alternatives!




Textual Document
Pre-Processing
Toponym Mentions
Sentence Windows
Paragraph Windows
ELMo/BERT
ELMo/BERT
ELMo/BERT
Contextual Word Embedding Features
Bidirectional LSTM
Bidirectional LSTM
Bidirectional LSTM
Max Pooling
Max Pooling
Max Pooling
Concatenation
+
Use Representations to Predict Toponym Locations
x y z
Number of Classes
Centroid Coordinates for HEALPix Classes
Interpolation
Output
Geographic Coordinates
Great Circle Distance
HEALPix Classes + Coordinates
Output
HEALPix Classification
Cross Entropy
Dense Layer

Overview

- Introduction
- Previous approaches
- A neural method for toponym resolution
- **Results**
- Extensions to base model
- Conclusions and future work

Datasets and metrics

- Several datasets have been used in previous studies
  - War of the Rebellion (WOTR)
  - Local-global-lexicon (LGL)
  - SpatialML
  - Alternative: Dataset from recent SemEval competition
- **Compute distance between estimated coordinates and ground-truth**

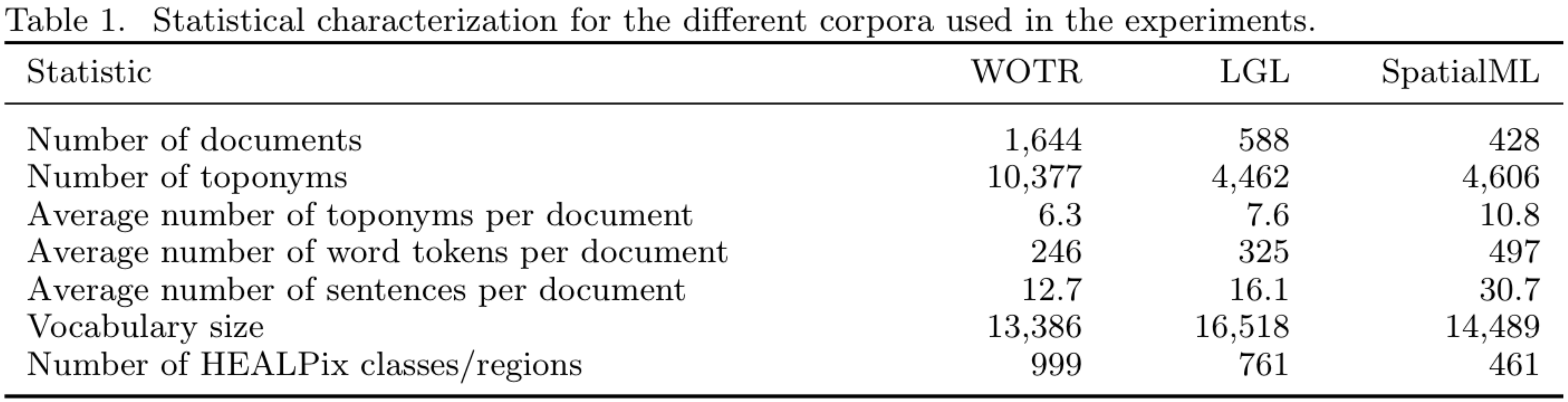


Table 1. Statistical characterization for the different corpora used in the experiments.

Statistic	WOTR	LGL	SpatialML
Number of documents	1,644	588	428
Number of toponyms	10,377	4,462	4,606
Average number of toponyms per document	6.3	7.6	10.8
Average number of word tokens per document	246	325	497
Average number of sentences per document	12.7	16.1	30.7
Vocabulary size	13,386	16,518	14,489
Number of HEALPix classes/regions	999	761	461

Table 2. Experimental results obtained with the base ELMo models.

Dataset	Mean dist. (km)	Median dist.(km)	Accuracy@161 km (%)
WOTR corpus			
TopoCluster (DeLozier *et al.* 2016)	604	—	57.0
TopoClusterGaz (DeLozier *et al.* 2016)	468	—	72.0
GeoSem (Ardanuy and Sporleder 2017)	445	—	68.0
Our Neural Model	**164**	11.48	**81.5**
LGL corpus			
GeoTxt (Gritta *et al.* 2018a)	1400	—	68.0
CamCoder (Gritta *et al.* 2018a)	700	—	76.0
TopoCluster (DeLozier *et al.* 2015)	1029	28.00	69.0
TopoClusterGaz (DeLozier *et al.* 2016)	1228	**0.00**	71.4
Learning to Rank (Santos *et al.* 2015)	742	2.79	—
Our Neural Model	**237**	12.24	**86.1**
SpatialML corpus			
Learning to Rank (Santos *et al.* 2015)	**140**	28.71	—
Our Neural Model	395	**9.08**	87.4

Overview

- Introduction
- Previous approaches
- A neural method for toponym resolution
- Results
- **Extensions to base model**
- Conclusions and future work

Possible extensions

- Mining additional training data from Wikipedia
  - Wikipedia pages associated to geospatial coordinates
  - Explore links from Wikipedia text into pages with geospatial coordinates
  - Easy to collect large amounts of (multilingual) data

- **Leverage geophysical terrain properties**
  - **Intuition:** Context of place reference often discusses these properties
  - Terrain elevation, land coverage, area occupied by water, etc.
  - Collect a summary value (e.g., average) for each HEALPix region
  - Same approach used to integrate proximity towards geospatial coordinates

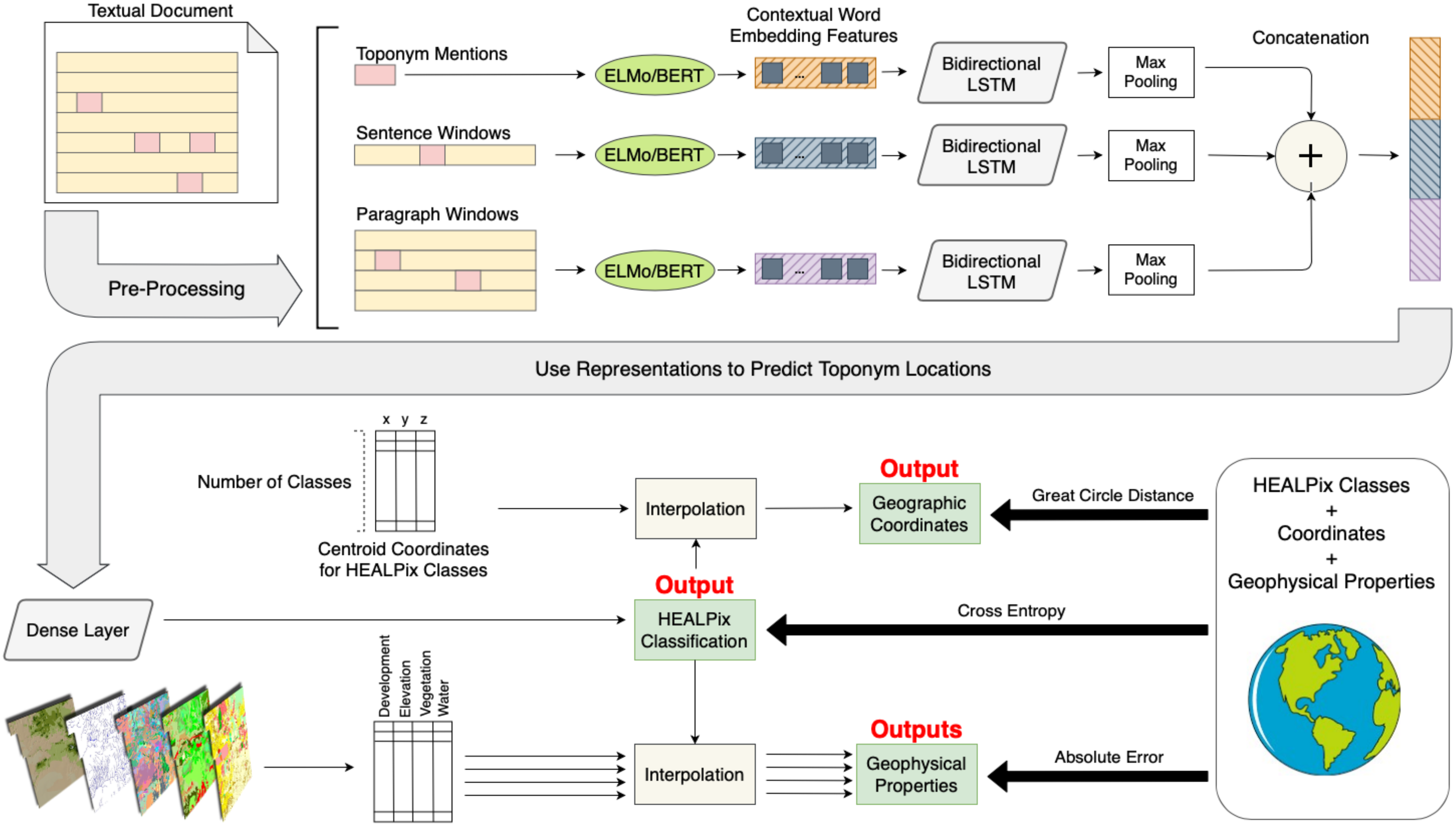

Textual Document
Pre-Processing
Toponym Mentions
Sentence Windows
Paragraph Windows
ELMo/BERT
ELMo/BERT
ELMo/BERT
Contextual Word Embedding Features
...
Bidirectional LSTM
Bidirectional LSTM
Bidirectional LSTM
Max Pooling
Max Pooling
Max Pooling
Concatenation
+
Use Representations to Predict Toponym Locations
x y z
Number of Classes
Centroid Coordinates for HEALPix Classes
Interpolation
Output
Geographic Coordinates
Great Circle Distance
HEALPix Classes + Coordinates + Geophysical Properties
Output
HEALPix Classification
Cross Entropy
Dense Layer
Development
Elevation
Vegetation
Water
Interpolation
Outputs
Geophysical Properties
Absolute Error

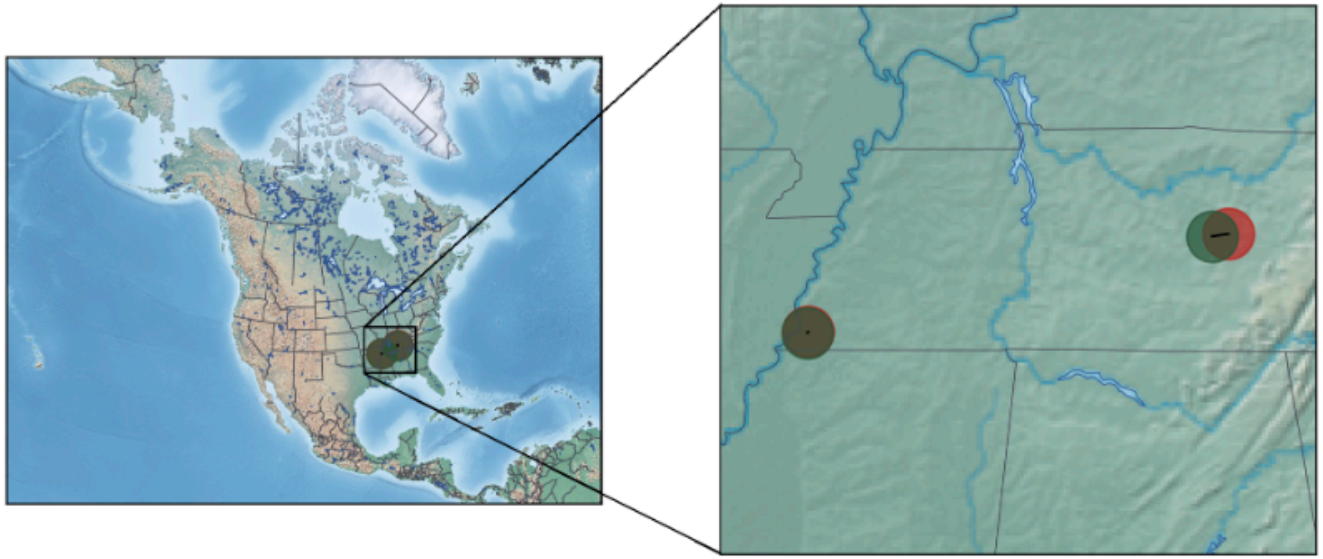
Table 3. Experimental results obtained with different modeling alternatives.

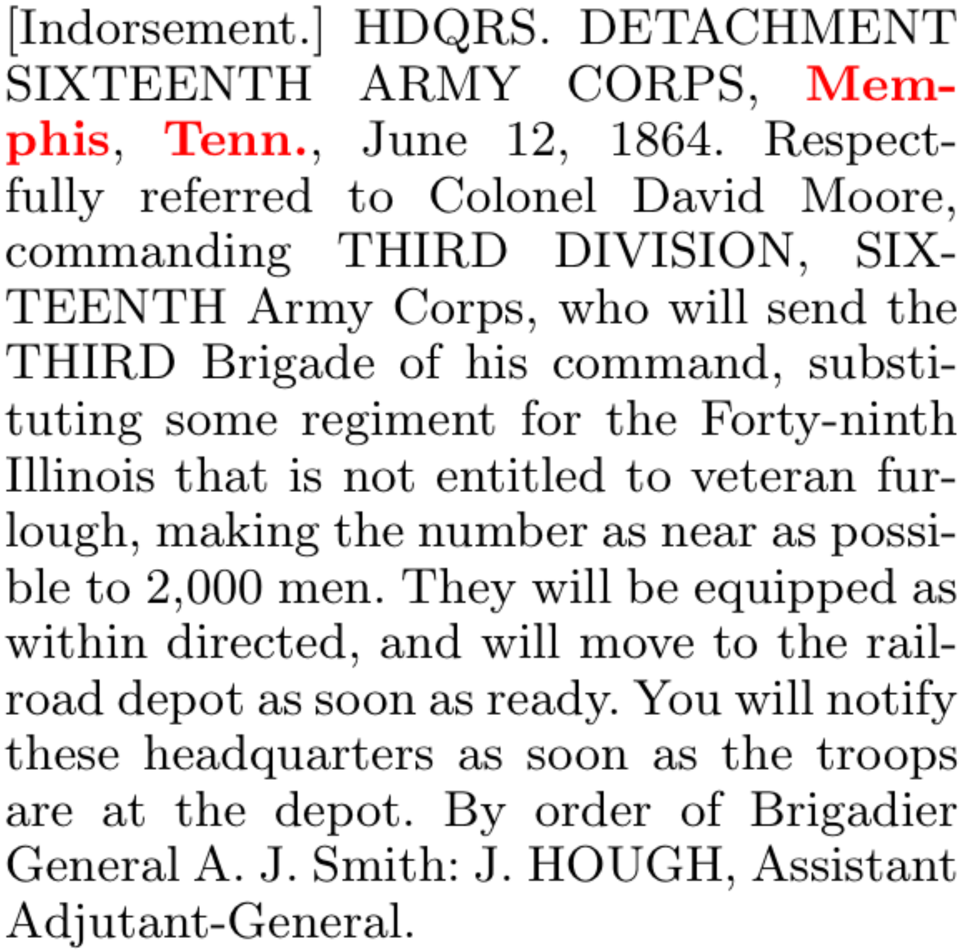
Model and dataset	Mean distance (km)	Median distance (km)	Accuracy@161 km (%)
WOTR corpus			
ELMo	164	11.48	81.5
ELMo + Wikipedia	158	11.28	82.4
ELMo + Geophysical	166	11.35	81.9
BERT	117	**10.99**	**87.3**
BERT + Wikipedia	122	11.04	86.4
BERT + Geophysical	**114**	**10.99**	**87.3**
LGL corpus			
ELMo	237	12.24	86.1
ELMo + Wikipedia	304	12.16	87.4
ELMo + Geophysical	282	12.24	87.7
BERT	**193**	11.81	90.1
BERT + Wikipedia	226	**11.51**	**90.6**
BERT + Geophysical	216	12.24	87.9
SpatialML corpus			
ELMo	395	**9.08**	87.4
ELMo + Wikipedia	364	**9.08**	88.5
ELMo + Geophysical	387	**9.08**	87.4
BERT	363	**9.08**	89.2
BERT + Wikipedia	**205**	**9.08**	**92.4**
BERT + Geophysical	339	**9.08**	89.4

Table 4. Examples of toponyms, taken from the different corpora, with low/high prediction errors.

Corpus	Lowest distance errors (km)	Highest distance errors (km)
WOTR	(0.63) Mexico (1.00) Resaca (1.09) Owen's Big Lake	(3104.59) Fort Welles (3141.29) Washington (3682.01) Astoria
LGL	(1.21) W.Va. (1.36) Butler County (1.51) Manchester	(8854.04) Ohioans (9225.86) North America (9596.54) Nigeria
SpatialML	(0.45) Tokyo (2.38) Lusaka (2.44) English	(9687.43) Capital (10818.50) Omaha (13140.64) Atlantic City

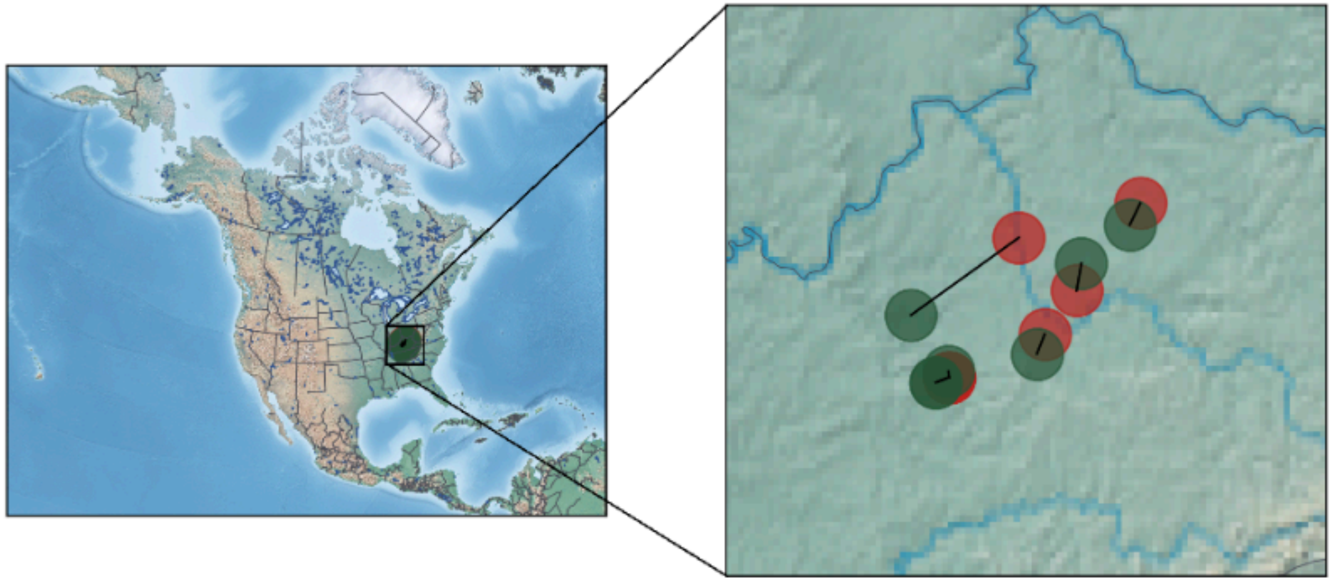
Predicted versus ground-truth coordinates (1)





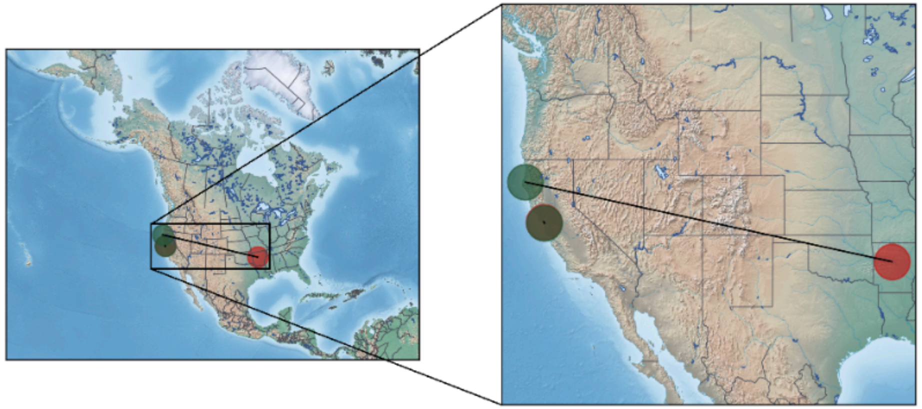
[Indorsement.] HDQRS. DETACHMENT SIXTEENTH ARMY CORPS, **Memphis**, **Tenn.**, June 12, 1864. Respectfully referred to Colonel David Moore, commanding THIRD DIVISION, SIXTEENTH Army Corps, who will send the THIRD Brigade of his command, substituting some regiment for the Forty-ninth Illinois that is not entitled to veteran furlough, making the number as near as possible to 2,000 men. They will be equipped as within directed, and will move to the railroad depot as soon as ready. You will notify these headquarters as soon as the troops are at the depot. By order of Brigadier General A. J. Smith: J. HOUGH, Assistant Adjutant-General.

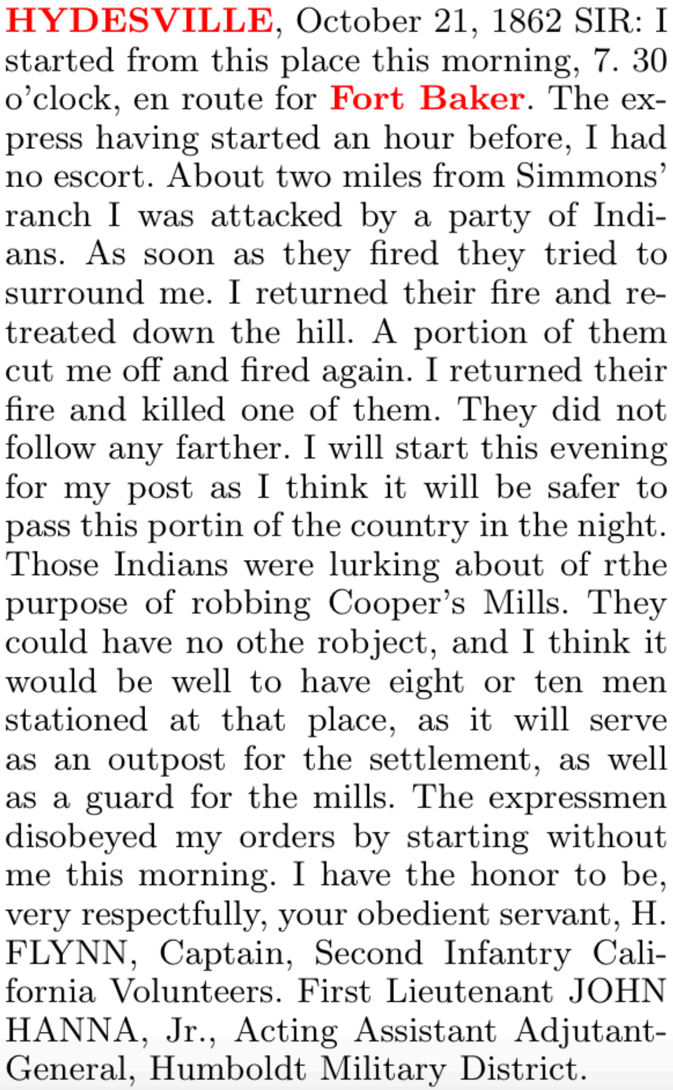
Predicted versus ground-truth coordinates (2)



LEXINGTON, **KY.**, June 11, 1864–11 p. m. Colonel J. W. WEATHERFORD, **Lebanon**, **Ky.** Have just received dispatch from General Burbridge at **Paris**. He says direct Colonel Weatherford to closely watch in the direction of **Bardstown** and **Danville**, and if any part of the enemy's force appears in that region to attack and destroy it. J. BATES DICKSON, Captain and Assistant Adjutant-General.

Predicted versus ground-truth coordiantes (3)





HYDESVILLE, October 21, 1862 SIR: I started from this place this morning, 7. 30 o'clock, en route for **Fort Baker**. The express having started an hour before, I had no escort. About two miles from Simmons' ranch I was attacked by a party of Indians. As soon as they fired they tried to surround me. I returned their fire and retreated down the hill. A portion of them cut me off and fired again. I returned their fire and killed one of them. They did not follow any farther. I will start this evening for my post as I think it will be safer to pass this portin of the country in the night. Those Indians were lurking about of rthe purpose of robbing Cooper's Mills. They could have no othe robject, and I think it would be well to have eight or ten men stationed at that place, as it will serve as an outpost for the settlement, as well as a guard for the mills. The expressmen disobeyed my orders by starting without me this morning. I have the honor to be, very respectfully, your obedient servant, H. FLYNN, Captain, Second Infantry California Volunteers. First Lieutenant JOHN HANNA, Jr., Acting Assistant Adjutant-General, Humboldt Military District.

Overview

- Introduction
- Previous approaches
- A neural method for toponym resolution
- Results
- Extensions to base model
- **Conclusions and future work**

Conclusions

- **What do you think of these ideas?**
  - Matching with gazetteers versus directly predicting geospatial coordinates?
    - Depends on the application!
  - Interest in using geophysical properties?
    - Opens many possibilities in terms of multimodal data analysis!

Many possibilities for improvement

- Further tests with extension related to geophysical properties
- Fine-tune large Transformer models (*instead of using them as feature extractors*)
- Replace regular grid supporting the classification task
  - Use non-uniform approach based on quadtrees
  - Use multiple overlapping partitions (and multiple classification outputs)
- Predict regions instead of coodinates for point locations
- Interpretability and better handling vague/approximate place references