Entity and Relation Extraction

Sandeep Soni

11/13/2023

Words

They can fish

Sequence Labeling

- Given a textual sequence, categorize each token in the sequence

$$f([x_1, x_2, \ldots, x_n]) \rightarrow [y_1, y_2, \ldots, y_n]$$

- Examples:
  - Part of speech tagging
  - Named entity recognition

Named entities

- Named entities are proper nouns that reference an entity that can be grounded
- Named entity recognition (NER) is the task of identifying spans of text that are named entities and categorizing them

BIO NOTATION

B	Beginning of a named entity
I	Inside of a named entity
O	Outside of the named entity

B-ORG	O	O	O	B-DATE	I-DATE	O	B-PER	I-PER	O	B-LOC
Apple	was	founded	on	April	1976,	by	Steve	Jobs	in	California

- If there are n categories, then there are 2n+1 BIO tags

NER TAGS

Person	Proper nouns refer to people or characters (e.g., Harry Potter, Sandeep Soni)	PER
Location	Refer to natural geographical regions (e.g., mountains) or structures built by humans (e.g., building)	LOC
GPE	Geo political entities such as cities, states, countries,etc (e.g., London)	GPE
Organization	Organization names, company names, sports teams, etc (e.g., The United Nations)	ORG

SPACY TAGS

CARDINAL	LAW	PERCENT
DATE	LOC	PERSON
EVENT	MONEY	PRODUCT
FAC	NORP	QUANTITY
GPE	ORDINAL	TIME
LANGUAGE	ORG	WORK_OF_ART

```
PERSON:      People, including fictional.
NORP:        Nationalities or religious or political groups.
FAC:         Buildings, airports, highways, bridges, etc.
ORG:         Companies, agencies, institutions, etc.
GPE:         Countries, cities, states.
LOC:         Non-GPE locations, mountain ranges, bodies of water.
PRODUCT:     Objects, vehicles, foods, etc. (Not services.)
EVENT:       Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART: Titles of books, songs, etc.
LAW:         Named documents made into laws.
LANGUAGE:    Any named language.
DATE:        Absolute or relative dates or periods.
TIME:        Times smaller than a day.
PERCENT:     Percentage, including ”%“.
MONEY:       Monetary values, including unit.
QUANTITY:    Measurements, as of weight or distance.
ORDINAL:     “first”, “second”, etc.
CARDINAL:    Numerals that do not fall under another type.
```

https://www.dataknowsall.com/ner.html

Fine grained NER

- Expanding the number of categories for named entities gives us more ways to distinguish between them

person	doctor	**organization**	terrorist_organization
actor	engineer	airline	government_agency
architect	monarch	company	government
artist	musician	educational_institution	political_party
athlete	politician	fraternity_sorority	educational_department
author	religious_leader	sports_league	military
coach	soldier	sports_team	news_agency
director	terrorist		

location	body_of_water	**product**	camera	**art**	written_work
city	island	engine	mobile_phone	film	newspaper
country	mountain	airplane	computer	play	music
county	glacier	car	software		
province	astral_body	ship	game	**event**	military_conflict
railway	cemetery	spacecraft	instrument	attack	natural_disaster
road	park	train	weapon	election	sports_event
bridge				protest	terrorist_attack

building	time	chemical_thing	website
airport	color	biological_thing	broadcast_network
dam	award	medical_treatment	broadcast_program
hospital	educational_degree	disease	tv_channel
hotel	title	symptom	currency
library	law	drug	stock_exchange
power_station	ethnicity	body_part	algorithm
restaurant	language	living_thing	programming_language
sports_facility	religion	animal	transit_system
theater	god	food	transit_line

Ling and Weld 2012

Domain specific NER

- The categories are highly dependent on the domain of interest
- Of course, the black King is out in the open, but the white Bishop is hemmed in by the pawns.

Tag	Meaning
Hu	Human
Tu	Turn
Po	Position
Pi	Piece
Ps	Piece specifier
Mc	Move compliment
Pa	Piece attribute
Pq	Piece quantity
Re	Region
Ph	Phase
St	Strategy
Ca	Castle
Me	Move eval.
Mn	Move name
Ee	Eval. element
Ev	Evaluation
Ti	Time
Ac	Player action
Ap	Piece action
Ao	Other action
Ot	Other notion

Tomori et. al (2016) in ACL

Nested NER

- Typically NER assumes all categories are in a flat structure.
  - E.g., [Emory University]$_{\text{ORG}}$
- But many named entities have a nested structure
  - E.g., [[Emory]$_{\text{PER}}$ University]$_{\text{ORG}}$

Named entity recognition

- Many different ways to identify named entities
  - Max Entropy Models (e.g., Multinomial Logistic Regression)
  - Max Entropy Markov Models
  - Recurrent Neural networks
  - Transformer models

Named entity linking

- Recognize spans that are named entities
- Also, resolve the named entities to their references
- Usually, by linking them to some external database

Paris is the capital of France

[Paris]$_{\text{LOC}}$ is the capital of [France]$_{\text{LOC}}$

Paris, Arkansas

Paris, France

[Paris]$_{LOC}$ is the capital of [France]$_{LOC}$

https://en.wikipedia.org/wiki/Paris

[Paris] is the capital of [France]

https://en.wikipedia.org/wiki/France

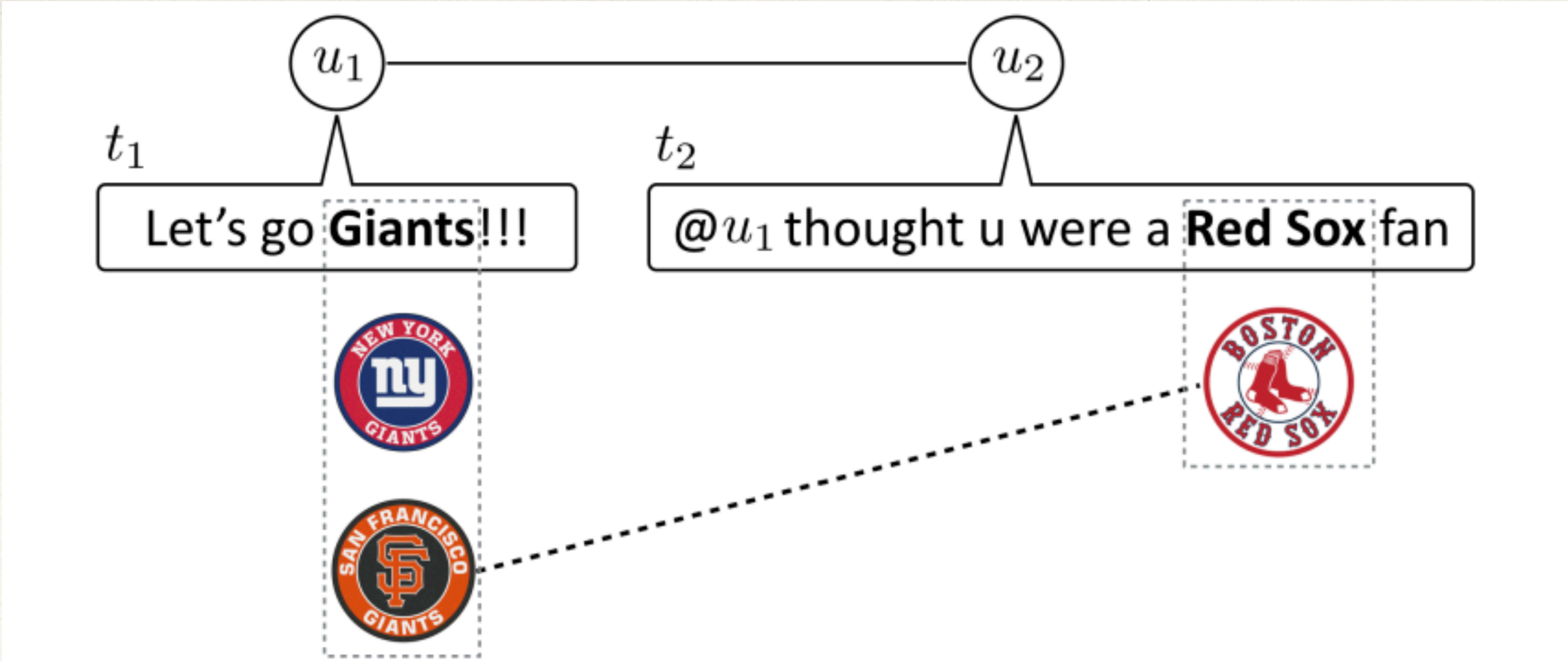


Figure 1: Illustration on leveraging social relations for entity disambiguation. Socially connected users u_1 and u_2 tend to talk about similar entities (baseball in the example).

Yi Yang, Ming-Wei Chang, and Jacob Eisenstein. 2016. Toward Socially-Infused Information Extraction: Embedding Authors, Mentions, and Entities. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1452–1461, Austin, Texas. Association for Computational Linguistics.

Toponym resolution

- Named entity linking but restricted to geographic entities.
- The objective is usually to link a span of text that marks a geographic entity to geographical coordinates on a map

Toponym resolution

High-speed rail between Toronto and London by 2025, premier says

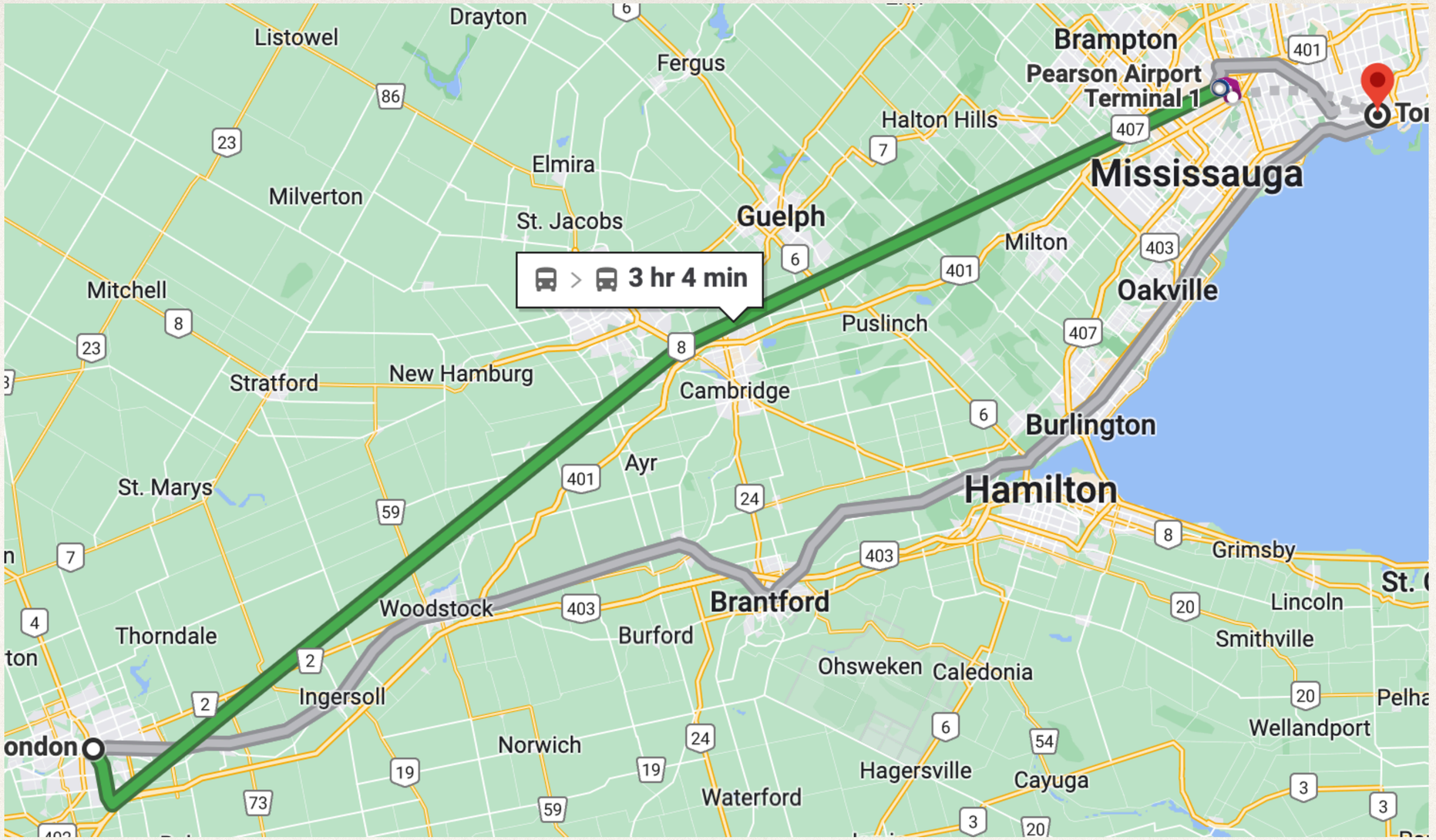
Wynne: 'Time to take a serious look at an idea that's been around for decades.'

CBC News · Posted: May 19, 2017 9:22 AM EDT | Last Updated: May 19, 2017

https://www.cbc.ca/news/canada/kitchener-waterloo/wynne-ontario-high-speed-rail-report-1.4123183

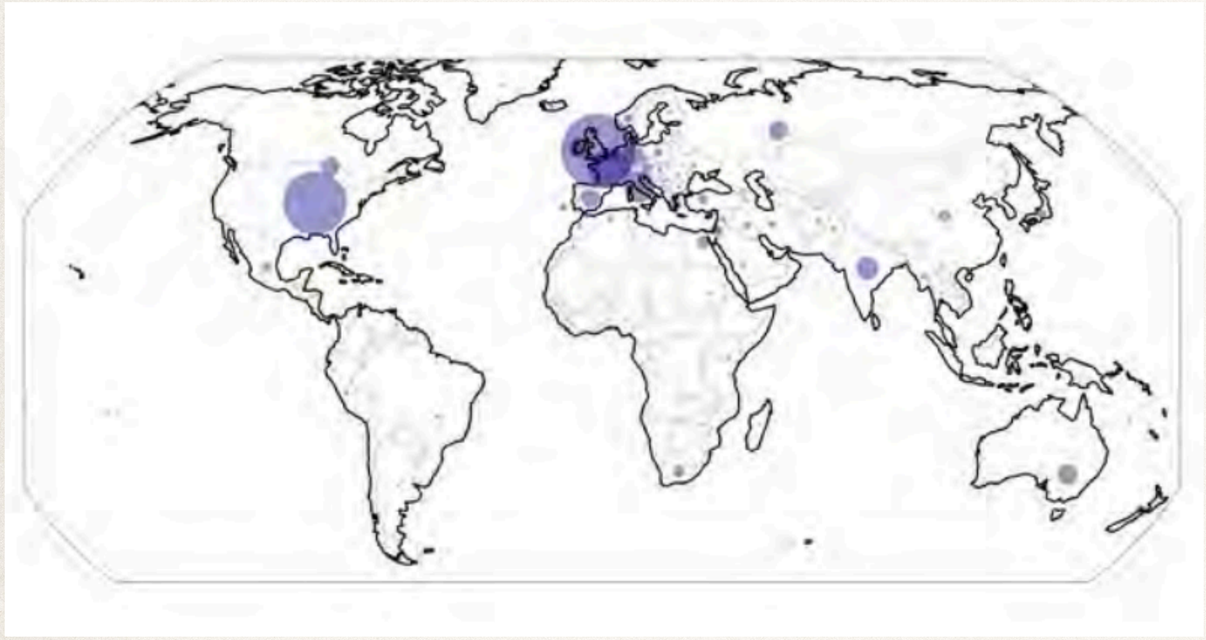
Toponym resolution

Can there be a high speed rail between London, UK to Toronto, Canada?



In general, we can resolve the place names by linking the mentions to a latitude and longitude

What can geographical named entities tell us about British Fiction?

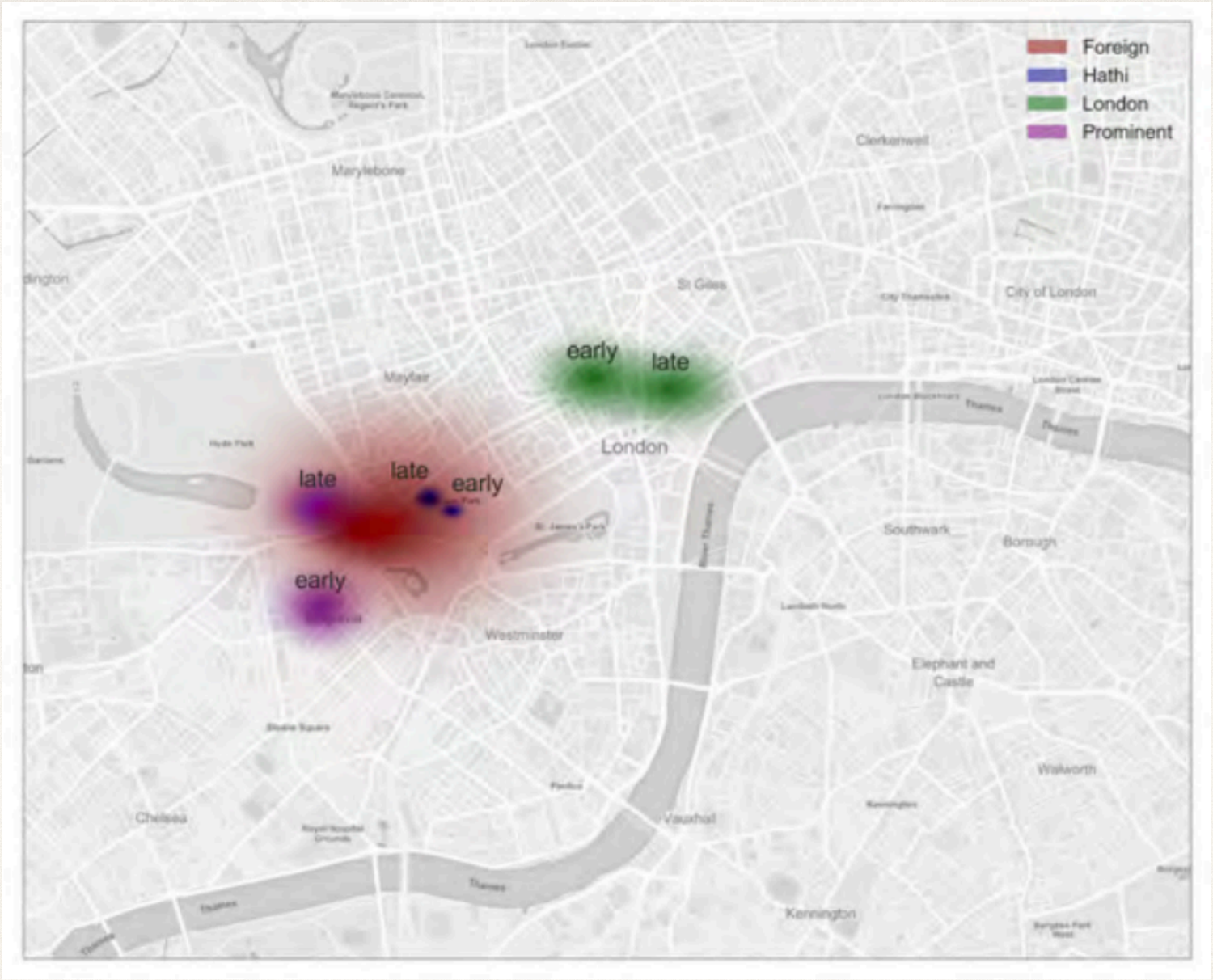


The distribution of global geographical attention in two large book corpora for British fiction

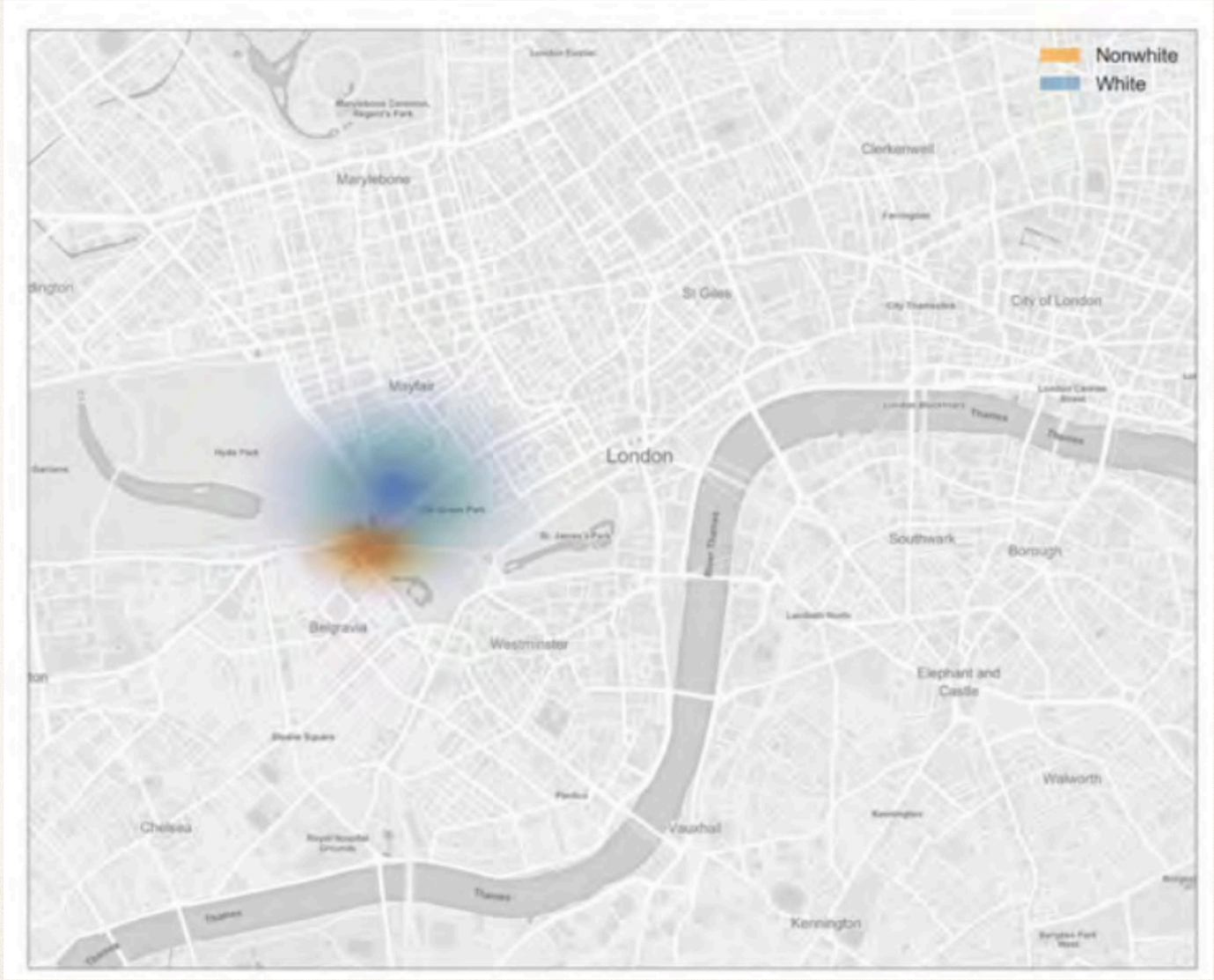
Evans, Elizabeth, and Matthew Wilkens. "Nation, ethnicity, and the geography of british fiction, 1880-1940." *Journal of Cultural Analytics* 3.2 (2018).

The distribution of attention over places in London in two large book corpora for British fiction

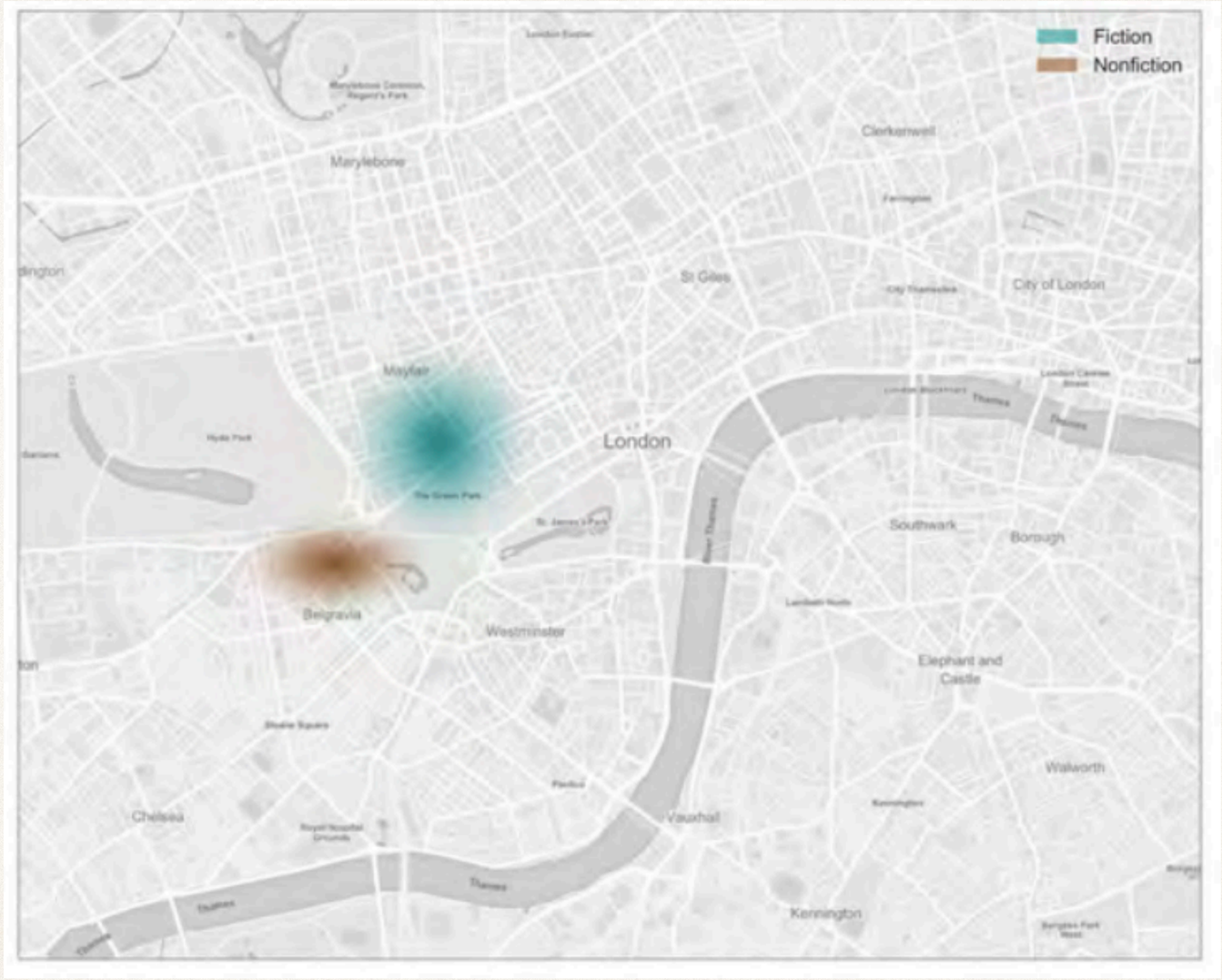
Evans, Elizabeth, and Matthew Wilkens. "Nation, ethnicity, and the geography of british fiction, 1880-1940." *Journal of Cultural Analytics* 3.2 (2018).



Evans, Elizabeth, and Matthew Wilkens. "Nation, ethnicity, and the geography of british fiction, 1880-1940." *Journal of Cultural Analytics* 3.2 (2018).



Evans, Elizabeth, and Matthew Wilkens. "Nation, ethnicity, and the geography of british fiction, 1880-1940." *Journal of Cultural Analytics* 3.2 (2018).



Evans, Elizabeth, and Matthew Wilkens. "Nation, ethnicity, and the geography of british fiction, 1880-1940." *Journal of Cultural Analytics* 3.2 (2018).

Relation extraction

- Named entities mentioned in texts can be used to construct a knowledge base
- Two entities are connected to each other if they have a relationship

My [apartment]$_{\text{LOC}}$ has a large [kitchen]$_{\text{LOC}}$

My [apartment]$_{LOC}$ has a large [kitchen]$_{LOC}$

My [apartment]$_{LOC}$ has a large [kitchen]$_{LOC}$

COMPONENT_OF (“Apartment”, “Kitchen”)

[Attention head]$_{\text{MODEL}}$ is at the heart of [Transformers]$_{\text{MODEL}}$

COMPONENT_OF (“Transformers”, “Attention head”)

[Attention head]$_{\text{MODEL}}$ is at the heart of [Transformers]$_{\text{MODEL}}$

COMPONENT_OF (“Transformers”, “Attention head”)

[Barack Obama]$_{PER}$ was born in [Hawaii]$_{LOC}$

ENTITY_ORIGIN (“Barack Obama”, “Hawaii”)

[Barack Obama]$_{PER}$ was born in [Hawaii]$_{GPE}$

ENTITY_ORIGIN (PER, GPE)

[Barack Obama]$_{PER}$ was born in [Hawaii]$_{GPE}$
He was the president of the [United States of America]$_{GPE}$
He is the husband of [Michelle Obama]$_{PER}$

ENT_ORIGIN	Barack Obama	Hawaii
PRESIDENT	Barack Obama	United States of America
SPOUSE	Barack Obama	Michell Obama

In class

- Sequence labeling demo
- Modify this notebook to find the distribution of named entities in a book from Project Gutenberg