LREC-COLING 2024

Automatic Construction of a Large-Scale Corpus for Geoparsing Using Wikipedia Hyperlinks

[1]Keyaki Ohno, [1]Hirotaka Kameko, [1]Keisuke Shirai, [2]Taichi Nishimura, [1]Shinsuke Mori

[1]Kyoto University [2]LY Corporation

Geoparsing and annotated datasets

Geoparsing = Geotagging + Geocoding

Melbourne is located within Ontario, Canada.

Geoparsing and annotated datasets

Geoparsing = Geotagging + Geocoding

Location Expression (Toponym) Recognition

Approach: Sequential Labeling
Dataset: NER datasets

Loc Loc Loc
Melbourne is located within Ontario, Canada.

Geoparsing and annotated datasets

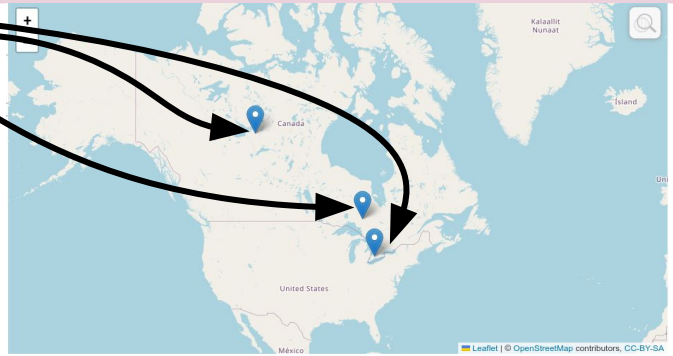
Geoparsing = Geotagging + Geocoding

Coordinates (latitude and longitude) Estimation

Approach 1: Gazetteer-based
Dataset: Gazetteer (e.g. GeoNames)

Approach 2: Machine learning-based
Dataset: annotated texts

Loc Loc Loc
Melbourne is located within Ontario, Canada.



Geoparsing and annotated datasets

Geoparsing = Geotagging + Geocoding

Location Expression (Toponym) Recognition

Approach: Sequential Labeling
Dataset: NER datasets

Coordinates (latitude and longitude) Estimation

Approach 1: Gazetteer-based
Dataset: Gazetteer (e.g. GeoNames)

Approach 2: Machine learning-based
Dataset: annotated texts

Need annotated dataset for training and evaluation

Human annotation
✔ Highly reliable
✖ High annotation cost → Small-scale

Automatic annotation (e.g. pattern match)
✔ Low annotation cost → Large-scale
✖ Low quality

Wikipedia Article and its Coordinates

Melbourne, Ontario (https://en.wikipedia.org/wiki/Melbourne,_Ontario)

Add languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Coordinates: 42°49′00″N 81°33′07″W

Melbourne is a small community located within Middlesex County, Ontario, Canada. It lies on the boundary between two municipalities, Strathroy-Caradoc and Southwest Middlesex. About half the population of Melbourne lives in each municipality.

The community was probably named for Melbourne, Victoria, Australia.[1]

Melbourne	
Community	
Country	Canada
Province	Ontario
County	Middlesex

Melbourne (https://en.wikipedia.org/wiki/Melbourne) 166 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Coordinates: 37°48′51″S 144°57′47″E

This article is about the Australian metropolitan area. For other uses, see Melbourne (disambiguation).

Melbourne (/ˈmɛlbərn/ *MEL-bərn*;[note 1] locally ['mælbən], Boonwurrung/Woiwurrung: *Narrm* or *Naarm*[9][10]) is the capital of the state of Victoria and the second-most populous city in Australia (behind Sydney), although the most populous by contiguous urban area.[11] Its name generally refers to a 9,993 km² (3,858 sq mi) metropolitan area also known as **Greater Melbourne**.[12]

Melbourne
Naarm (Woiwurrung)
Naarm (Boonwurrung)
Victoria

- Articles
- Coordinates
- Hyperlinks

are edited by human

Overview of WHLL

WHLL: **W**ikipedia **H**yperlink-based **L**ocation **L**inking method

0. Download CirrusSearch dump and HTML dumps
1. Make a dictionary to convert from title to coordinates
2. List up articles about locations
3. Extract spans matching the title
4. Extract spans with hyperlinks referring to locations

1. Create (Title → Coordinates) Dictionary

Melbourne	-37.81417	144.96306
Melbourne, Ontario	42.81667	-81.55194
Middlesex County, Ontario	43.00000	-81.50000
Ontario	49.25000	-84.50000
Canada	60.00000	-110.00000
Victoria (Australia)	-37.00000	144.00000

(Title, latitude, longitude)

from CirrusSearch dump

2. Search for Articles with Coordinates

Title: Melbourne, Ontario / Coordinates: (42.81667, -81.55194)

Text: Melbourne is a small community located within

Middlesex County,

Ontario, Canada.

Melbourne	-37.81417	144.96306
Melbourne, Ontario	42.81667	-81.55194
Middlesex County, Ontario	43.00000	-81.50000
Ontario	49.25000	-84.50000
Canada	60.00000	-110.00000
Victoria (Australia)	-37.00000	144.00000

The community was probably named for Melbourne,

Victoria, Australia.

3. Search for Spans matching the Title

Title: Melbourne, Ontario / Coordinates: (42.81667, -81.55194)

Text: Melbourne (42.81667, −81.55194) is a small community located within

y,_Ontario”>Middlesex County,

, Canada.

complete title
or
rule-based ordinary names
(commas or parentheses)

The community was probably named for Melbourne,

Victoria, Australia.

Melbourne	-37.81417	144.96306
Melbourne, Ontario	42.81667	-81.55194
Middlesex County, Ontario	43.00000	-81.50000
Ontario	49.25000	-84.50000
Canada	60.00000	-110.00000
Victoria (Australia)	-37.00000	144.00000

3. Search for Spans matching the Title

Title: Melbourne, Ontario / Coordinates: (42.81667, -81.55194)

Text: Melbourne (42.81667, −81.55194) is a small community l

Middlesex County,

Ontario, Canada.

Skip linked spans

The community was probably named for Melbourne,

Victoria, Australia.

Melbourne	-37.81417	144.96306
Melbourne, Ontario	42.81667	-81.55194
Middlesex County, Ontario	43.00000	-81.50000
Ontario	49.25000	-84.50000
Canada	60.00000	-110.00000
Victoria (Australia)	-37.00000	144.00000

4. Get Coordinates of referred Articles

Title: Melbourne, Ontario / Coordinates: (42.81667, -81.55194)

Text: Melbourne is a small community located within
(42.81667, −81.55194)

Middlesex County,
(43.00000, −81.50000)

Ontario, Canada.
(49.25000, −84.50000) (60.00000, −110.00000)

The community was probably named for Melbourne,
(−37.81417, 144.96306)

Victoria, Australia.
(−37.00000, 144.00000)

Melbourne	-37.81417	144.96306
Melbourne, Ontario	42.81667	-81.55194
Middlesex County, Ontario	43.00000	-81.50000
Ontario	49.25000	-84.50000
Canada	60.00000	-110.00000
Victoria (Australia)	-37.00000	144.00000

4. Get Coordinates of referred Articles

Title: Melbourne, Ontario / Coordinates: (42.81667, -81.55194)

Text: Melbourne (42.81667, −81.55194) is a small community located within

Middlesex County (43.00000, −81.50000),

Ontario (49.25000, −84.50000), Canada (60.00000, −110.00000).

The community was probably named for Melbourne (−37.81417, 144.96306),

Victoria (−37.00000, 144.00000), Australia.

"Australia" is not linked to any articles

Melbourne	-37.81417	144.96306
Melbourne, Ontario	42.81667	-81.55194
Middlesex County, Ontario	43.00000	-81.50000
Ontario	49.25000	-84.50000
Canada	60.00000	-110.00000
Victoria (Australia)	-37.00000	144.00000

Statistics of WHLL dataset

Language : en.wikipedia
CirrusSearch dump: 2023-07-10
HTML dump : 2023-07-01

Total	
#articles	1,315,117
#sentences	23,334,035
#location expressions	14,726,908
Ambiguous	45.6%
Ambiguous & Recessive	9.9%
#unique LEs	1,571,291
(Ambiguous	8.1%

Per article	
#sentences	17.7
#tokens	420.1
#location expressions	11.3
#unique LEs	7.8

Ambiguous:= expressions that have the same string but are tied to different points
Recessive:= ambiguous expressions and do not refer to the most freq. coordinates

Advantage of WHLL Datasets

- **Large-scale** and **general** domain
  - → entire Wikipedia articles
- Reliable annotation with small cost
  - → data source (texts, hyperlinks, and coordinates) is edited by human
- Ambiguous expressions (for eval. disambiguation)
  - → same as general texts
- Publicly available (for reproducibility)
  - → same as Wikipedia copyright policy (CC BY-SA 4.0 and GFDL)
    (We published our WHLL script under MIT license.)

Advantage of WHLL Datasets

- Large-scale and general domain
  → entire Wikipedia articles
- **Reliable** annotation with **small cost**
  → data source (texts, hyperlinks, and coordinates) is edited by human
- Ambiguous expressions (for eval. disambiguation)
  → same as general texts
- Publicly available (for reproducibility)
  → same as Wikipedia copyright policy (CC BY-SA 4.0 and GFDL)
  (We published our WHLL script under MIT license.)

Advantage of WHLL Datasets

- Large-scale and general domain
  - → entire Wikipedia articles
- Reliable annotation with small cost
  - → data source (texts, hyperlinks, and coordinates) is edited by human
- **Ambiguous expressions** (for eval. disambiguation)
  - → same as general texts
- Publicly available (for reproducibility)
  - → same as Wikipedia copyright policy (CC BY-SA 4.0 and GFDL)
    (We published our WHLL script under MIT license.)

Advantage of WHLL Datasets

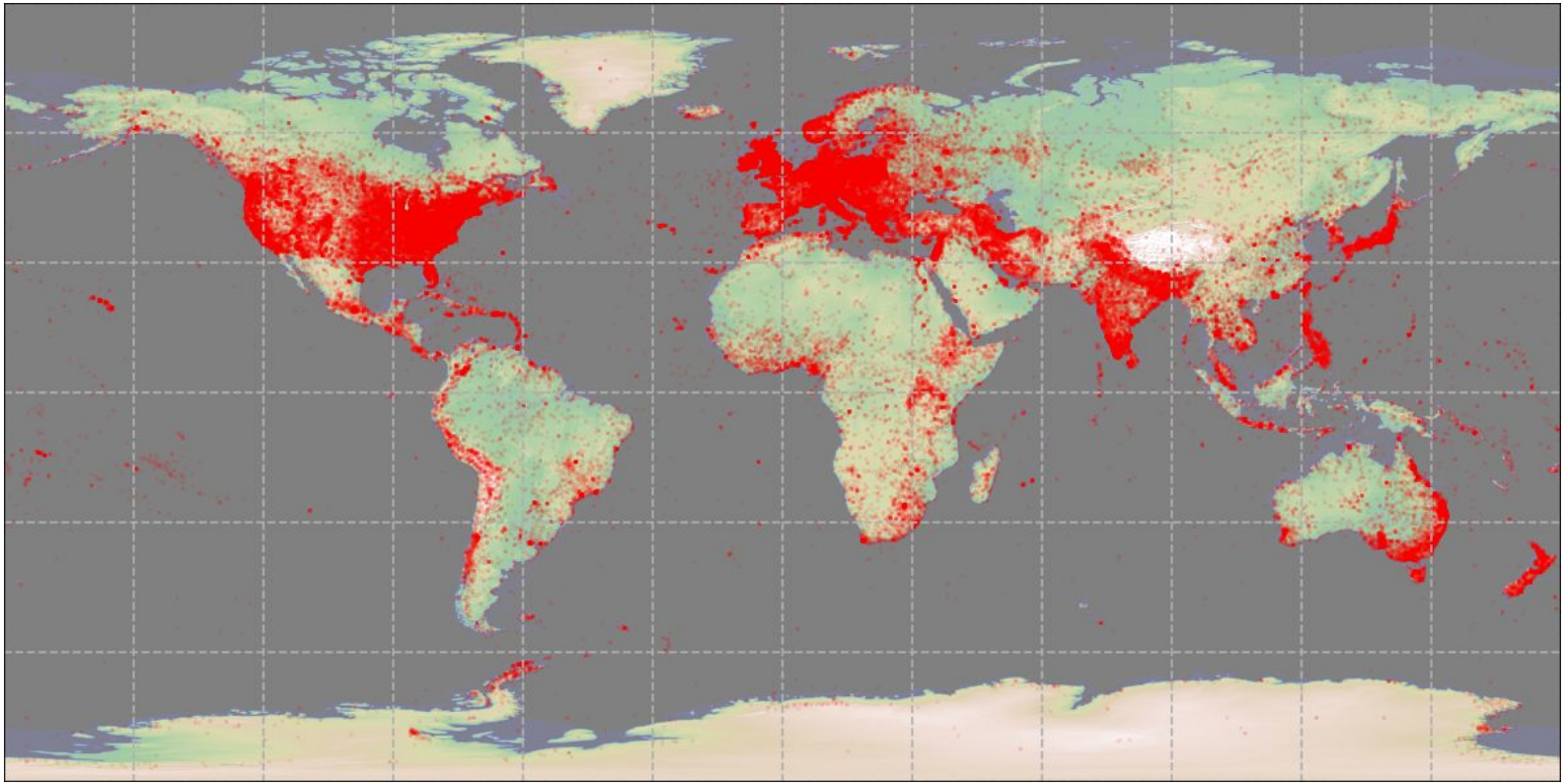
- Large-scale and general domain
  → entire Wikipedia articles
- Reliable annotation with small cost
  → data source (texts, hyperlinks, and coordinates) is edited by human
- Ambiguous expressions (for eval. disambiguation)
  → same as general texts
- **Publicly available** (for reproducibility)
  → same as Wikipedia copyright policy (CC BY-SA 4.0 and GFDL)
  (We published our WHLL script under MIT license.)

Publicly Available Geocoding Datasets (en)

Dataset	#Articles
LGL (Lieberman et al., 2010)	588
WikToR (Gritta et al., 2018)	5,000
GeoVirus (Gritta et al., 2018)	229
SemEval-2019 Task 12 (Weissenbacher et al., 2019)	150
TR-News (Kamalloo et al., 2018)	118
GeoWebNews (Gritta et al., 2020)	200
ours (WHLL en)	1,315,117

Large-scale annotated dataset
without any human effort just for WHLL

Coordinates appear 10 or more times



WHLL for Various Languages

WHLL can create datasets in the various languages in Wikipedia

Notation: WHLL–{edition_code}–CS{timestamp of CirrusSearch dump}.HTML{timestamp of HTML dump}

*#lines of HTML dumps

Dataset	#Articles	#Articles in Wikipedia*	#Location expressions
WHLL-en-CS20230710.HTML20230701	1,315,117	6,875,034	14,726,908
WHLL-ja-CS20240304.HTML20240301	200,906	1,439,102	4,151,205
WHLL-it-CS20240422.HTML20240420	349,124	1,904,627	2,539,082

Project Page

Our project page

(Codes and Datasets)

http://www.lsta.media.kyoto-u.ac.jp/resource/data/WHLL/home-e.html

GitHub repository

(Codes)

https://github.com/ku-lsta/WHLL