

Automatic Nested Spatial Entity and Spatial Relation Extraction From Text for Knowledge Graph Creation: A Baseline Approach and a Benchmark Dataset

Helen Mair RAWSTHORNE¹,

Nathalie ABADIE¹, Eric KERGOSIEN², Cécile DUCHÊNE¹, Eric SAUX³

¹LASTIG, Univ Gustave Eiffel, IGN-ENSG, France; ²GERiico, Université de Lille, France; ³IRENAV, École navale, France

13 november 2023 | Hamburg, Germany

7th ACM SIGSPATIAL International Workshop on Geospatial Humanities

Why extract spatial entities and spatial relations from text?

2

- Textual sources
 - Unstructured spatial knowledge
 - Current or historical
 - Examples: travel guides, historical documents, social media posts, etc.
- Spatial knowledge
 - Alternative names
 - Additional features
 - Unknown spatial entities
 - Geographic location information

Why store spatial knowledge in a knowledge graph?

- Heterogeneous spatial knowledge
 - Different authors
 - Different points of view
 - Different levels of detail
 - Vagueness
- Geospatial knowledge graphs
 - Queries
 - Spatial reasoning
 - Links to other sources

Why extract *nested* spatial entities from text?

- Flat spatial entity extraction
 - Port of Hamburg: **geographic name**
- Nested spatial entity extraction
 - Port of Hamburg: **geographic name**
 - Port: **geographic feature**
 - Hamburg: **name**

entity class assigned
automatically

Port of Hamburg

indication of
geographic location

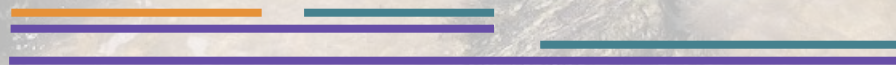
Why extract *nested* spatial entities from text?

- Flat spatial entity extraction
 - Robben Island Lighthouse: **geographic name**
- Nested spatial entity extraction
 - Robben Island: **geographic name**
 - Island: **geographic feature**
 - Robben: **name**
 - Robben Island Lighthouse: **geographic name**
 - Lighthouse: **geographic feature**

Robben Island Lighthouse

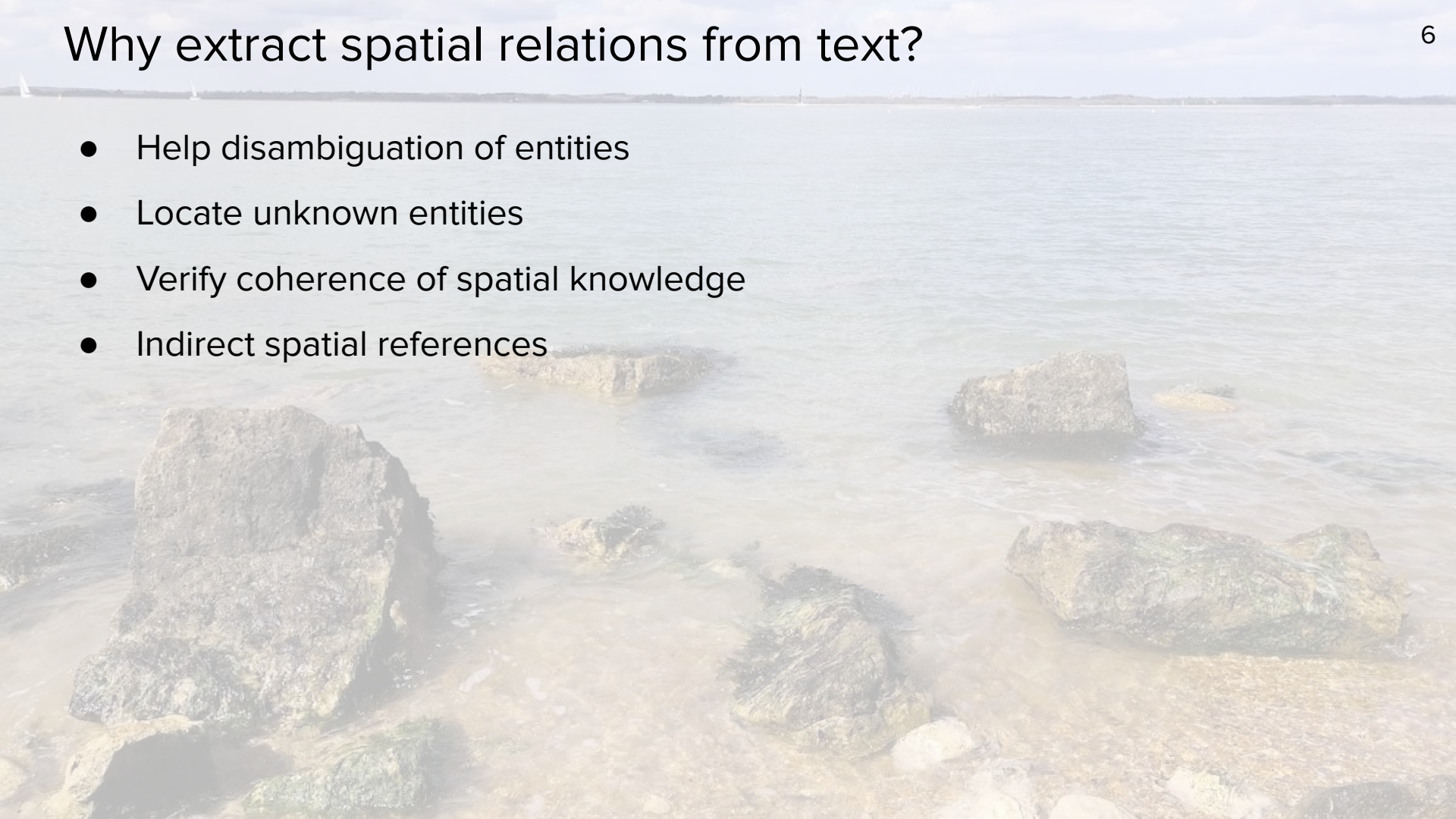
entity class assigned
automatically

indication of
geographic location



Why extract spatial relations from text?

- Help disambiguation of entities
- Locate unknown entities
- Verify coherence of spatial knowledge
- Indirect spatial references



- Structure geographic information from text as knowledge graph
 - Define ontology
 - Extract geographic information
 - Structure and disambiguate
- Application to the *Instructions nautiques*
 - Coastal navigation instructions
 - Coastal environment descriptions

25 À 1,9 M au SSW du phare d'Alger, un monument (36° 44,7' N — 3° 04,2' E) haut de 80 m est remarquable. Le monument et les pylônes environnants portent des feux d'obstacle aérien. À 1 M au SE du monument, le petit dôme blanc de l'ancien séminaire de **Kouba** (36° 44,1' N — 3° 05,3' E) est visible de toute la baie.

31 Le tronçon Nord de la jetée détachée est remarquable par les grands réservoirs qu'il porte.

37



- **Princeton University Relation Extraction system (PURE)** [Zhong and Chen 2021]
 - Prepare annotated training dataset
 - Flat entities
 - Binary relations
 - Train two pretrained deep language models
 - Flat entity extraction
 - Binary relation extraction
- Our approach: PURE with **nested spatial entities** and binary **spatial relations**
 - Context window size?
 - Monolingual vs multilingual models?

Dataset Preparation

- *Instruction nautiques* extracts
 - Worldwide coverage
 - ~101,000 tokens
 - ~17,000 entity labels
 - ~3,000 relation labels
- Manual annotation
 - *brat* rapid annotation tool

- Nested spatial entity annotations
 - **geographic name**
 - **geographic feature**
 - **name**
- Binary spatial relation annotations
 - Topological: “is off the coast of”, “is marked by”, “is an element of”
 - Absolute directional: 16 cardinal directions

Un port de pêche est établi à 5,7 M à l'ENE de Ras Magroua .

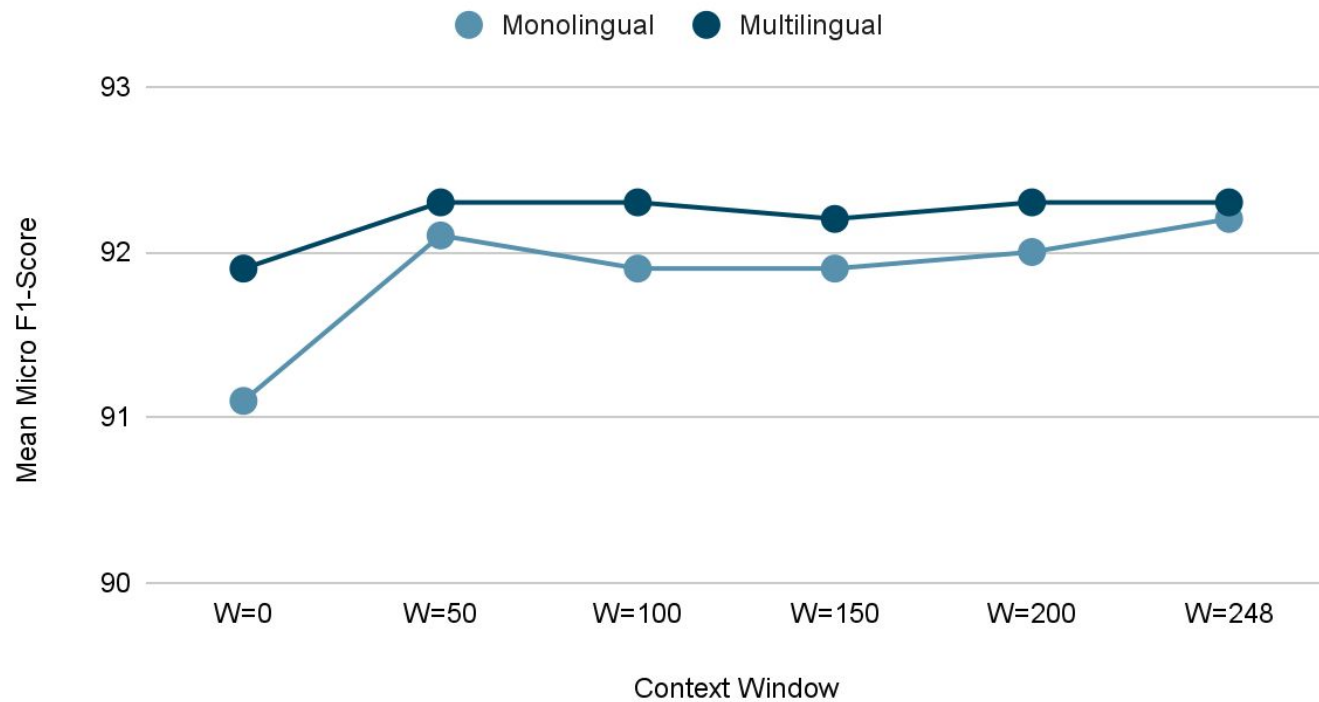
is ENE of

- Pretrained models
 - Monolingual: *bert-base-french-europeana-cased*
 - Multilingual: *bert-base-multilingual-cased*
- Context windows
 - Entity model: 0, 50, 100, 150, 200, 248
 - Relation model: 0, 50, 100

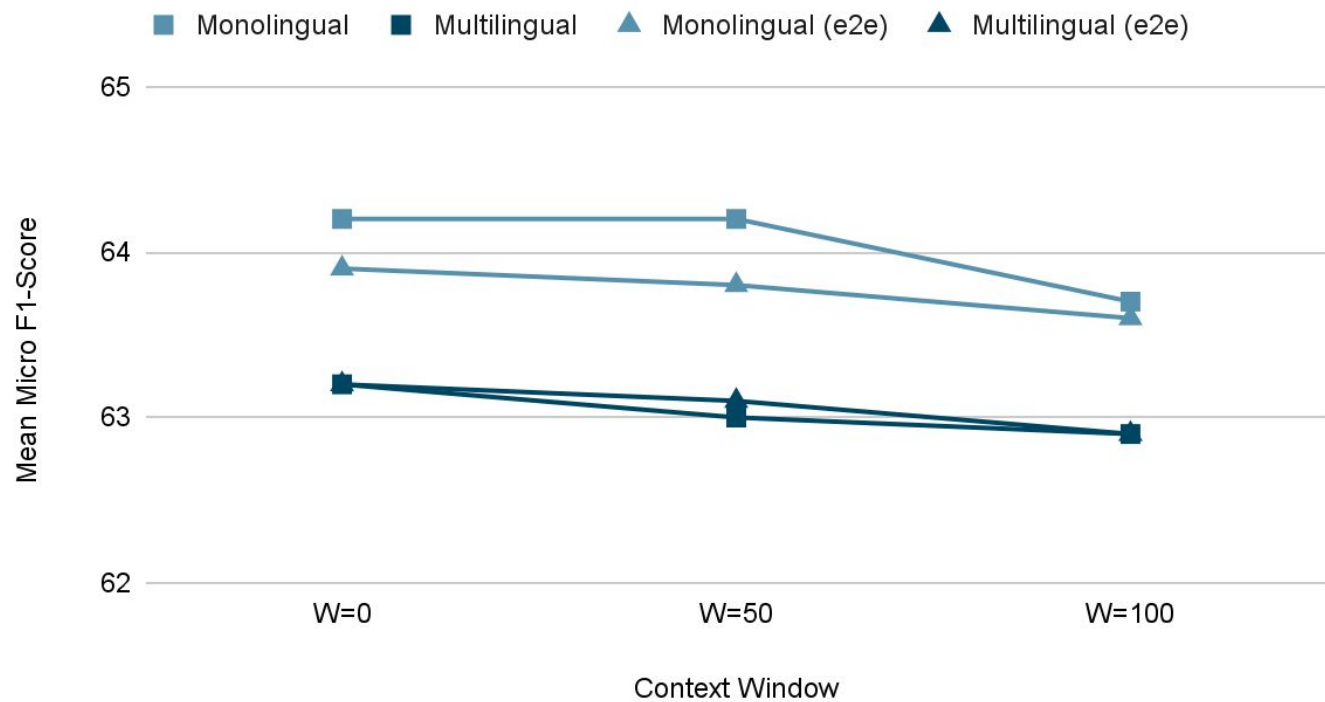


Hyperparameter	Entity Model	Relation Model
learning rate	1e-5	2e-5
task learning rate	5e-4	-
train batch size	16	32
training epochs	100	10

Entity Model



Relation Model



	F1-Score	
	Monolingual	Multilingual
Entity	92.2	92.3
Relation	64.2	63.2
Relation (e2e)	63.9	63.2

Results: Monolingual Example

1 01 4.6.3.



2 De Ras Magroua à Ténès

3 01 4.6.3.1.

4 Généralités



5 07 La côte de Ras Magroua à Ténès s'étend sur environ 28 M de l'WSW à l'ENE.

6 Elle est presque continûment rocheuse et escarpée.



7 Les dangers côtiers sont peu nombreux et, sauf à proximité à l'Est de Ras Magroua (§ 4.6.3.2.), l'isobathe de 20 m suit le rivage, à une distance comprise entre 0,3 et 0,5 M.

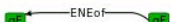
8 01 4.6.3.2.



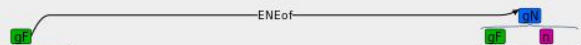
9 De Ras Magroua au-delà du phare de Nadji



10 07 À 7,5 M à l'ENE de Ras Kramis, Ras Magroua (cap Magroua) [36° 22,0' N — 0° 48,5' E] est abrupt.



11 À 0,7 M à l'ENE de ce cap se trouve une épave.



12 13 Un port de pêche (36° 24,1' N — 0° 54,6' E) est établi à 5,7 M à l'ENE de Ras Magroua.

13 19 À 2 M plus au Nord, Hadjrat Nadji (îlot Colombi) est un îlot rocheux, haut de 28 m, à 0,3 M au NW du rivage.

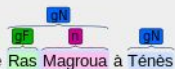
14 Il porte une balise (3 m) lumineuse (36° 26,2' N — 0° 55,1' E).

15 À 1,1 M au SSE, un piton conique, haut de 289 m, surmonté d'un signal (36° 25,6' N — 0° 55,8' E), est très reconnaissable.

16 25 À 1,1 M à l'ENE de Hadjrat Nadji, le phare de Nadji, tour carrée rouge à lanterne verte, haute de 29 m, se dresse près de la côte.

Results: Multilingual Example

1 01 4.6.3.



2 De Ras Magroua à Ténès

3 01 4.6.3.1.

4 Généralités



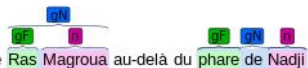
5 07 La côte de Ras Magroua à Ténès s'étend sur environ 28 M de l'WSW à l'ENE.

6 Elle est presque continûment rocheuse et escarpée.



7 Les dangers côtiers sont peu nombreux et, sauf à proximité à l'Est de Ras Magroua (§ 4.6.3.2.), l'isobathe de 20 m suit le rivage, à une distance comprise entre 0,3 et 0,5 M.

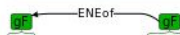
8 01 4.6.3.2.



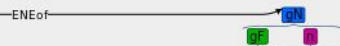
9 De Ras Magroua au-delà du phare de Nadji



10 07 À 7,5 M à l'ENE de Ras Kramis, Ras Magroua (cap Magroua) [36° 22,0' N — 0° 48,5' E] est abrupt.



11 À 0,7 M à l'ENE de ce cap se trouve une épave.



12 13 Un port de pêche (36° 24,1' N — 0° 54,6' E) est établi à 5,7 M à l'ENE de Ras Magroua.



13 19 À 2 M plus au Nord, Hadjrat Nadji (îlot Colombi) est un îlot rocheux, haut de 28 m, à 0,3 M au NW du rivage.

14 Il porte une balise (3 m) lumineuse (36° 26,2' N — 0° 55,1' E).



15 À 1,1 M au SSE, un piton conique, haut de 289 m, surmonté d'un signal (36° 25,6' N — 0° 55,8' E), est très reconnaissable.



16 25 À 1,1 M à l'ENE de Hadjrat Nadji, le phare de Nadji, tour carrée rouge à lanterne verte, haute de 29 m, se dresse près de la côte.

- PURE [Zhong and Chen 2021] → spatial entities, spatial relations
 - Nested entity extraction approach
- New French-language dataset
 - <https://github.com/umrlastig/atlantis-dataset> (from 16/11)
 - <https://www.kaggle.com/competitions/defi-textmine-2024/>
- Benchmark results

- Reducing annotation time
 - Smaller dataset + unsupervised learning
 - Machine-assisted annotation
 - Synthetic data
- Extending extraction scope
 - N-ary relations
 - Distance spatial relations
 - Intrinsic and relative directional spatial relations
 - Entity characteristics

Thank you!

Helen Mair RAWSTHORNE

helen.rawsthorne@gmail.com