Comparison of named entity recognition methods for geographical information retrieval

Hedi Zeghidi

Internship tutor: Ludovic Moncla

Referent teacher: Marc Sebban

04 July 2024

LABEX ASLAN
UNIVERSITÉ DE LYON

Université
Jean Monnet
Saint-Étienne

Summary

Comparison of named entity recognition methods

- Introduction, Team and Project
- Dataset Description
- Method comparison
- Experiment
- Conclusion, Reflection and Perspectives

Introduction, Team and Project

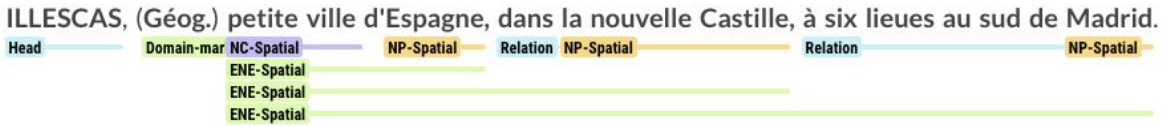
Comparison of named entity recognition methods

- Work at LIRIS laboratory, in the DM2L Team, within the GEODE project
- Conducting a state-of-the-art review of methodologies for the identification of named entities (NP), nominal entities (NC), nested entities (ENE) and spatial relationships (Relation)
- Create a hybrid method for the recognition and classification of named entities and spatial relationship

Dataset Description

Dictionary composed of 2200 entries from “French Encyclopédie ou dictionnaire raisonné des sciences des arts et des métiers par une société de gens de lettres” (1751-1772)

- NC-Spatial: a common noun that identifies a spatial entity (nominal spatial entity)
- NC-Person: a common noun that identifies a person (nominal person entity)
- NP-Spatial: a proper noun identifying the name of a place (spatial named entities)
- NP-Person: a proper noun identifying the name of a person (person named entities)
- NP-Misc: a proper noun identifying entities not classified as spatial or person
- ENE-Spatial: nested spatial entity
- ENE-Person: nested person entity
- ENE-Misc: nested named entity not classified as spatial or person
- Relation: spatial relation
- Latlong: geographic coordinates
- Head: entry name
- Domain-Mark: words indicating the knowledge domain



	Train	*Validation*	*Test*
Paragraphs	1,8	200	200
Tokens	132,398	14,959	13,881
NC-Spatial	3,252	358	355
NP-Spatial	4,707	464	519
ENE-Spatial	3,033	326	334
Relation	2,093	219	226
Latlong	553	66	72
NC-Person	1,378	132	133
NP-Person	1,599	170	150
ENE-Person	492	49	57
NP-Misc	948	108	96
ENE-Misc	255	31	22
Head	1,261	142	153
Domain-Mark	1,069	122	133

Method comparison

Models

- CRF Model NER
- Joint-Label CNN Model (nested entities of all levels)
- Joint-Label Bi-LSTM Model (nested entities of all levels)
- SPAN BERT (nested entities of all levels)
- GPT and LLMs

CRF Model NER

Train with a Gridsearch to tune hyperparameters for L1&L2 regularization

	base	base+POS	base+POS+DEP	Support
Domain-mark	98.2	98.6	99.0	392
Head	87.1	87.9	87.7	254
NC-Person	60.9	61.7	66.0	225
NC-Spatial	90.9	91.3	89.4	592
NP-Misc	60.3	64.1	63.8	175
NP-Person	75.9	75.3	77.4	203
NP-Spatial	89.8	90.2	91.1	718
Relation	92.8	92.7	91.0	452
Micro Avg	87.1	87.6	87.4	3011
Macro Avg	82.0	82.7	83.2	3011
Weighted Avg	86.5	87.0	87.1	3011

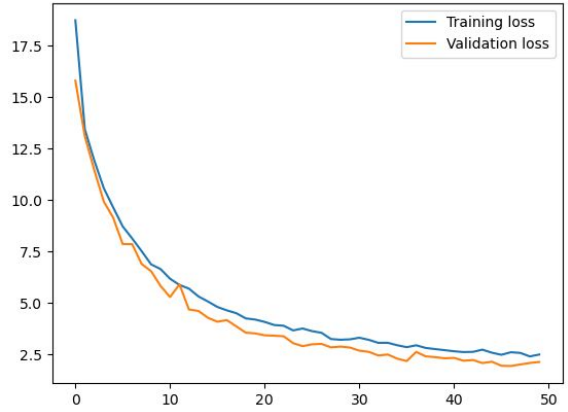
Features

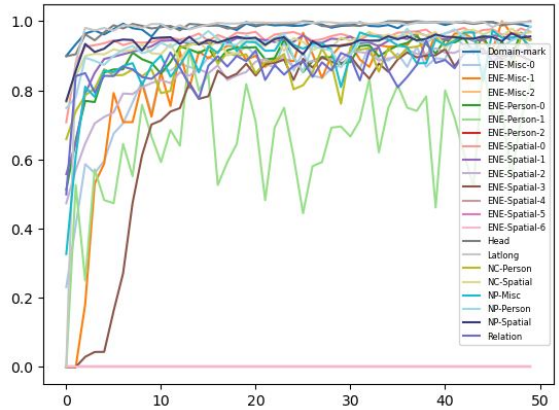
- Token: Word form
- lower: Lowercase word form
- isdigit: True if word is a number else False
- isupper: True if word is all capital letter else False
- ispunct: True if word is punctuation else False
- isstop: True if word is empty else False
- len: Number of characters composing the word
- shape: Shape of the word
- pos: Grammar part of the sentence (NOUN, VERB, etc)
- dep: Syntactic role of the word

	Domain mark	*Head*	*Relation*	*NC-Person*	*NC-Spatial*	*NP-Misc*	*NP-Person*	*NP-Spatial*
Top 5 best Features	'Token:.'	'prev_Token:*'	'lower:midi'	'lower:pape'	'lower:royaume'	'shape:dddd'	'prev_Token:Hunauld'	'Token:ltalie'
	'lower:.'	'prev_lower:*'	'Token:dessous'	'lower:roi'	'lower:fleuve'	'Token:persan'	'prev_lower:hunauld'	'lower:ltalie'
	'Token:terme'	'Token:)'	'lower:dessous'	'Token:Mans'	'lower:comté'	'lower:persan'	'lower:juifs'	'prev_lower:palus'
	'lower:terme'	'lower:)'	'Token:Midi'	'lower:mans'	'Token:île'	'prev_Token:xiij'	'Token:bazanés'	'prev_Token:palus'
	'Token:Géograph'	'isupper'	'next_Token:petits'	'lower:président'	'lower:île'	'prev_lower:xiij'	'lower:bazanés'	'lower:indes'
Top 5 worst Features	'shape:Xxxxxxxxxx'	'prev_lower:)'	'next_lower:se'	'shape:Xxxxxxxxx'	'next_dep:obl:agent'	'prev_Token:v.'	'prev_shape:Xxxxxxxx'	'shape:xxx'
	'dep:appos'	'prev_shape:X.'	'next_Token:sur'	'isstop'	'Token:du'	'next_shape:dddd'	'shape:xxxxxx'	'next_dep:flat:name'
	'pos:DET'	'prev_pos:NOUN'	'next_lower:sur'	'pos:ADV'	'lower:du'	'pos:ADP'	'shape:xxxx'	'shape:xxxxxxx'
	'pos:PUNCT'	'prev_pos:PUNCT'	'shape:x.'	'shape:Xxxxx'	'next_dep:det'	'shape:dd'	'shape:xxxxxxxxxx'	'shape:xxxxxx'
	'isupper'	'prev_shape:X'	'pos:PROPN'	'next_dep:xcomp'	'isstop'	'isstop'	'shape:xxxxxxxxx'	'shape:xxxxxxxxxx'

Joint-Label CNN Model all levels

Training during 50 epochs with a learning rate 0.001

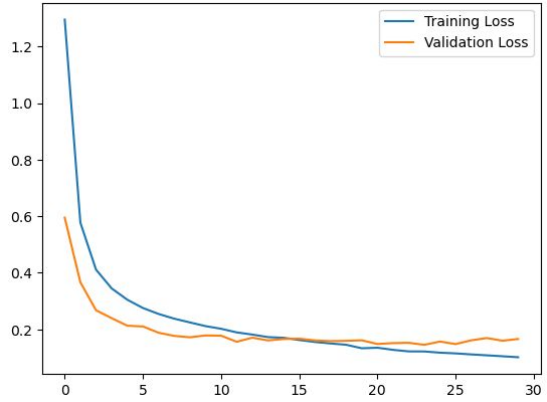


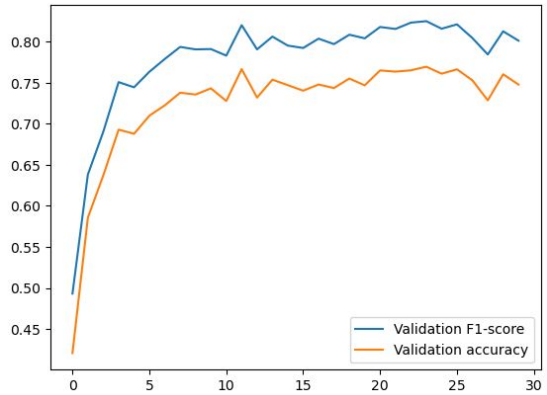


	Precision	*Recall*	*F-score*	*Support*		*Precision*	*Recall*	*F-score*	*Support*
Domain-mark	90.6	98.0	94.1	392	*ENE-Misc-2*	0.00	0.00	0.00	0
Head	91.2	85.8	88.4	254	*ENE-Person-0*	81.5	53.3	64.4	199
NC-Person	56.7	67.6	61.7	225	*ENE-Person-1*	50.0	4.8	8.7	21
NC-Spatial	91.4	83.1	87.1	592	*ENE-Person-2*	0.00	0.00	0.00	0
NP-Misc	43.3	66.3	52.4	175	*ENE-Spatial-0*	87.9	82.4	85.1	802
NP-Person	69.1	78.3	73.4	203	*ENE-Spatial-1*	77.3	61.8	68.7	685
NP-Spatial	90.1	77.0	83.0	718	*ENE-Spatial-2*	43.5	57.2	49.4	425
Relation	89.7	71.2	79.4	452	*ENE-Spatial-3*	24.2	9.1	13.3	175
Latlong	95.9	94.8	95.3	789	*ENE-Spatial-4*	21.7	12.5	15.9	40
ENE-Misc-0	27.1	28.4	27.7	81	*ENE-Spatial-5*	0.00	0.00	0.00	0
ENE-Misc-1	0.00	0.00	0.00	5	*ENE-Spatial-6*	0.00	0.00	0.00	0
Micro Avg	78.5	74.2	76.3	6223					
Macro Avg	51.4	46.9	47.6	6233					
Weighted Avg	79.4	74.2	76.1	6233					

Joint-Label Bi-LSTM Model all levels

Training during 30 epochs with a learning rate 0.200



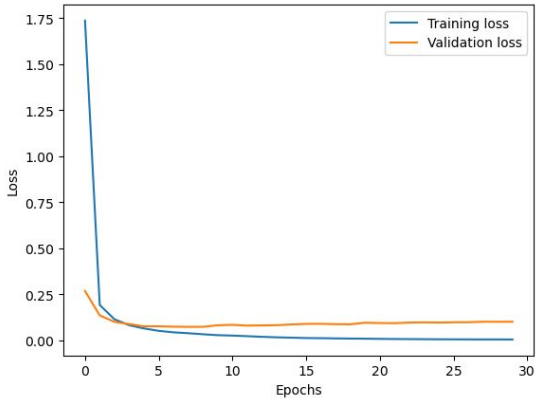


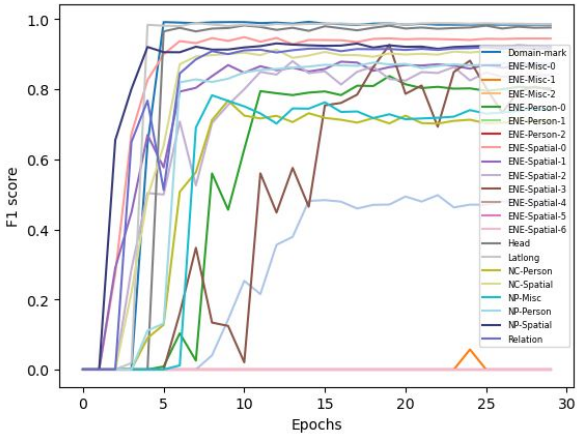
	Precision	*Recall*	*F-score*	*Support*
Domain-mark	99.2	99.0	99.1 +5.0	392
Head	96.0	94.9	95.4 +7.0	254
NC-Person	64.5	87.1	74.1 +12.4	225
NC-Spatial	90.3	95.6	92.9 +5.8	592
NP-Misc	73.1	76.0	74.5 +12.1	175
NP-Person	86.4	90.6	88.5 +15.1	203
NP-Spatial	95.7	95.3	95.5 +12.5	718
Relation	93.3	93.1	93.2 +13.8	452
Latlong	99.2	97.6	98.4 +3.1	789
ENE-Misc-0	37.0	42.0	39.3 +11.6	81
ENE-Misc-1	0.00	0.00	0.00 +0.0	5
Micro Avg	88.3	89.4	88.9 +12.6	6223
Macro Avg	61.7	59.2	59.0 +11.4	6233
Weighted Avg	89.2	89.4	89.0 +12.9	6233

	Precision	*Recall*	*F-score*	*Support*
ENE-Misc-2	0.00	0.00	0.00 +0.0	0
ENE-Person-0	90.9	70.4	79.3 +14.9	199
ENE-Person-1	100	19.0	32.0 +23.3	21
ENE-Person-2	0.00	0.00	0.00 +0.0	0
ENE-Spatial-0	93.3	89.9	91.6 +6.5	802
ENE-Spatial-1	91.2	83.2	87.0 +18.3	685
ENE-Spatial-2	76.9	91.1	83.4 +34.0	425
ENE-Spatial-3	69.3	76.0	72.5 +59.2	175
ENE-Spatial-4	1.8	2.5	2.1 -13.8	40
ENE-Spatial-5	0.00	0.00	0.00 +0.0	0
ENE-Spatial-6	0.00	0.00	0.00 +0.0	0

SPAN BERT all levels

Training during 30 epochs with a learning rate 0.0001





	Precision	*Recall*	*F-score*	*Support*		*Precision*	*Recall*	*F-score*	*Support*
Domain-mark	99.7	99.0	99.4 +0.3	392	*ENE-Misc-2*	0.00	0.00	0.00 +0.0	0
Head	97.3	98.0	97.6 +2.2	254	*ENE-Person-0*	88.3	79.9	83.9 +4.6	199
NC-Person	69.1	85.3	76.3 +2.2	225	*ENE-Person-1*	0.00	0.00	0.00 32.0	21
NC-Spatial	88.4	95.4	91.8 -1.1	592	*ENE-Person-2*	0.00	0.00	0.00 +0.0	0
NP-Misc	69.5	79.4	74.1 -0.4	175	*ENE-Spatial-0*	93.2	92.5	92.9 +1.3	802
NP-Person	87.6	86.7	87.1 -1.4	203	*ENE-Spatial-1*	84.7	85.4	85.0 2.0	685
NP-Spatial	97.0	94.4	95.7 +0.2	718	*ENE-Spatial-2*	76.1	94.6	84.4 +1.0	425
Relation	86.9	95.6	91.0 -1.2	452	*ENE-Spatial-3*	70.3	88.0	78.2 +5.7	175
Latlong	96.3	98.1	97.2 -1.2	789	*ENE-Spatial-4*	0.00	0.00	0.00 2.1	40
ENE-Misc-0	35.1	49.4	41.0 +1.7	81	*ENE-Spatial-5*	0.00	0.00	0.00 +0.0	0
ENE-Misc-1	0.00	0.00	0.00 +0.0	5	*ENE-Spatial-6*	0.00	0.00	0.00 +0.0	0
Micro Avg	87.4	91.0	89.2 +0.3	6223					
Macro Avg	56.3	60.1	58.0 -1.0	6233					
Weighted Avg	87.3	91.0	89.0 +0.0	6233					

Method comparison

	BASE+POS+DEP FEATURES CRF MODEL NER	*BASE BERT WITH ALL TAGS*	*SPAN BERT WITH ALL TAGS*	*SPACY SPANCAT*	*BI-LSTM MODEL WITH ALL TAGS*	*CNN Model WITH ALL TAGS*
Domain-mark	99.0	**99.9**	99.4	95.8	99.1	94.1
Head	87.7	97.5	**97.6**	45.1	95.4	88.4
Relation	91.0	91.4	91.0	52.5	**93.2**	79.4
Latlong	-	97.4	97.2	0.00	**98.4**	95.3
NC-Person	66.0	73.5	76.3	**78.0**	74.1	61.7
NC-Spatial	89.4	92.6	91.8	95.3	**92.9**	87.1
NP-Misc	63.8	71.5	74.1	71.9	**74.5**	52.4
NP-Person	77.4	87.3	87.1	**93.0**	88.5	73.4
NP-Spatial	91.1	95.4	**95.7**	95.4	95.5	83.0
ENE-Misc-0	-	**50.5**	41.0	0.00	39.3	27.7
ENE-Misc-1	-	0.00	0.00	0.00	0.00	0.00
ENE-Person-0	-	84.0	83.9	**88.2**	79.3	64.4
ENE-Person-1	-	**64.9**	0.00	41.0	32.0	8.7
ENE-Spatial-0	-	92.3	92.9	**94.1**	91.6	85.1
ENE-Spatial-1	-	86.0	85.0	**89.6**	87.0	68.7
ENE-Spatial-2	-	79.7	84.4	**87.8**	83.4	49.4
ENE-Spatial-3	-	67.7	**78.2**	78.0	72.5	13.3
ENE-Spatial-4	-	36.4	0.00	**51.6**	2.1	15.9

GPT and LLMs (Chat-GPT 3.5)

Presenting the task and formatting examples

```
You are an expert in Natural Language Processing. Your task is to identify common Named Entities (NER) in a given text.
The possible common Named Entities (NER) types are exclusively: (Domain-mark, Head, NC-Person, NC-Spatial, NP-Misc, NP-Person, NP-Spatial, Relation) and can be described as
:
1.Domain-mark: words indicating the knowledge domain (usually after the head and between parenthesis) such as 'Géog., Géog. mod., Géog. anc., Géogr., Géogr. mod., Marine., H
ist. nat., Gram., Géogr. anc., Jurisprud., Géog. anc. & mod., Gramm., Geog.'
2.Head: entry name at the beginning of the sentence and is almost always in uppercase such as 'Aire, Afrique, Aigle, ILLESCAS, MULHAUSEN, ADDA, SINTRA ou CINTRA, ACHSTEDE, o
u AKSTEDE, KEITH, CAÇERES, CARMAGNOLE, AGRIGNON, INSPRUCK'
3.NC-Person: a common noun that identifies a person such as 'M., roi, S., peuples, l'empereur, son fils, les habitans, prince, peuple, le roi, fils, le P., habitans'
4.NC-Spatial: a common noun that identifies a spatial entity including natural features such as 'ville, petite ville, la riviere, la mer, royaume, la province, capitale, la
ville, l'île, cette ville, pays, la côte, riviere'
5.NP-Misc: a proper noun identifying entities not classified as spatial or person such as 'l'Eglise, grec, 1707, russien, Glaciale, Noire, romain, la Croix, Russien, Parleme
nt, 1693, Sud, 1614'
6.NP-Person: a proper noun identifying the name of a person (person named entities) such as 'Ptolomée, Pline, Strabon, Euripide, les Romains, Pierre, Romains, les Anglois, T
urcs, Dieu, César, Antonin, les Espagnols'
7.NP-Spatial: a proper noun identifying the name of a place (spatial named entities) such as 'France, Allemagne, Italie, Espagne, Afrique, Asie, Paris, Naples, Angleterre, R
ome, Russie, la Chine, l'Amérique méridionale'
8.Relation: spatial relation such as 'dans, sur, au, en, entre, près de, se jette dans, proche, par, vers, près du, jusqu'à, à l'orient'.
9.Latlong: geographic coordinates such as 'Long. 31. 58. lat. 40. 55, Long. 10. 27. lat. 43. 30, Long. 28. 14. lat. 51. 13, Long. 14. 46. lat. 56. 20, Long. 12. 8. lat. 39.
15, Long. 25. 20. lat. 44. 43, Lat. 19. 40, Long. selon Harris, 29. 16. 15. lat. 47. 15, Long. 14. 28. lat : 53. 50, Long. 57. lat. 38. 35, Long. 22. 52. lat. 43. 32, Long.
11. 18. lat. 40. 41, Long. 27. 40. lat. 51. 50'.

Here are some examples:
EXAMPLE 1:
    INPUT:('PIRANO' ,0) (',' ,1) ('(' ,2) ('Géog' ,3) ('.' ,4) ('mod' ,5) ('.' ,6) (')' ,7) ('ville' ,8) ('d'' ,9) ('Italie' ,10) ('dans' ,11)
('l'' ,12) ('Istrie' ,13) (',' ,14) ('environ' ,15) ('à' ,16) ('14' ,17) ('milles' ,18) ('de' ,19) ('Capo' ,20) ('d'' ,21) ('Istria' ,22) (',' ,2
3) ('en' ,24) ('tirant' ,25) ('vers' ,26) ('le' ,27) ('midi' ,28) ('occidental' ,29) ('.' ,30) ('Elle' ,31) ('est' ,32) ('sur' ,33) ('une' ,34)
('petite' ,35) ('presqu'' ,36) ('île' ,37) ('formée' ,38) ('par' ,39) ('le' ,40) ('golfe' ,41) ('Largone' ,42) (',' ,43) ('&' ,44) ('celui' ,45)
('de' ,46) ('Trieste' ,47) ('.' ,48) ('Les' ,49) ('Vénitiens' ,50) ('en' ,51) ('sont' ,52) ('les' ,53) ('maîtres' ,54) ('depuis' ,55) ('1583' ,56)
('.' ,57) ('Long' ,58) ('.' ,59) ('31' ,60) ('.' ,61) ('46' ,62) ('.' ,63) ('lat' ,64) ('.' ,65) ('45' ,66) ('.' ,67) ('48' ,68) ('.' ,69)

    OUTPUT:[{'label': 'Head', 'text': 'PIRANO', 'start': 0, 'end': 0}, {'label': 'Domain-mark', 'text': 'Géog. mod.', 'start': 3, 'end': 6}, {'lab
el': 'NC-Spatial', 'text': 'ville', 'start': 8, 'end': 8}, {'label': 'NP-Spatial', 'text': 'Italie', 'start': 10, 'end': 10}, {'label': 'Relation'
, 'text': 'dans', 'start': 11, 'end': 11}, {'label': 'NP-Spatial', 'text': "l'Istrie", 'start': 12, 'end': 13}, {'label': 'Relation', 'text': 'env
iron à 14 milles de', 'start': 15, 'end': 19}, {'label': 'NP-Spatial', 'text': "Capo d'Istria", 'start': 20, 'end': 22}, {'label': 'Relation', 'te
xt': 'vers le midi occidental', 'start': 26, 'end': 29}, {'label': 'NC-Spatial', 'text': "une petite presqu'île", 'start': 34, 'end': 37}, {'labe
l': 'Relation', 'text': 'formée par', 'start': 38, 'end': 39}, {'label': 'NC-Spatial', 'text': 'le golfe', 'start': 40, 'end': 41}, {'label': 'NP-
Spatial', 'text': 'Largone', 'start': 42, 'end': 42}, {'label': 'NP-Spatial', 'text': 'Trieste', 'start': 47, 'end': 47}, {'label': 'NP-Person', '
text': 'Les Vénitiens', 'start': 49, 'end': 50}, {'label': 'NC-Person', 'text': 'les maîtres', 'start': 53, 'end': 54}, {'label': 'NP-Misc', 'text
': '1583', 'start': 56, 'end': 56}, {'label': 'Latlong', 'text': 'Long. 31. 46. lat. 45. 48', 'start': 58, 'end': 68}]
---
```

GPT and LLMs (Chat-GPT 3.5)

```
('HILPERHAUSEN' ,0) (',' ,1) ('(' ,2) ('Géog' ,3) ('.' ,4) (')' ,5) ('ville' ,6) ('d'' ,7) ('Allemagne' ,8) ('en' ,9) ('Franconie' ,10) (',' ,11) ('sur'
,12) ('la' ,13) ('Werra' ,14) (',' ,15) ('au' ,16) ('comté' ,17) ('de' ,18) ('Henneberg' ,19) (',' ,20) ('entre' ,21) ('Cobourg' ,22) ('&' ,23) ('Smalcal
de' ,24) (';' ,25) ('elle' ,26) ('appartient' ,27) ('à' ,28) ('une' ,29) ('branche' ,30) ('de' ,31) ('la' ,32) ('maison' ,33) ('de' ,34) ('Saxe' ,35)
('-' ,36) ('Gotha' ,37) ('.' ,38) ('Long' ,39) ('.' ,40) ('28' ,41) ('.' ,42) ('15' ,43) ('.' ,44) ('lat' ,45) ('.' ,46) ('50' ,47) ('.' ,48) ('35' ,49)
('.' ,50) ('(' ,51) ('D.' ,52) ('J.' ,53) (')' ,54)
```

```
{'entities': [{'text': 'HILPERHAUSEN', 'start': 0, 'end': 0, 'label': 'Head'},
  {'text': 'Géog.', 'start': 3, 'end': 4, 'label': 'Domain-mark'},
  {'text': 'ville', 'start': 6, 'end': 6, 'label': 'NC-Spatial'},
  {'text': 'Allemagne', 'start': 8, 'end': 8, 'label': 'NP-Spatial'},
  {'text': 'Franconie', 'start': 10, 'end': 10, 'label': 'NP-Spatial'},
  {'text': 'Werra', 'start': 14, 'end': 14, 'label': 'NC-Spatial'},
  {'text': 'comté', 'start': 17, 'end': 17, 'label': 'NC-Spatial'},
  {'text': 'Henneberg', 'start': 19, 'end': 19, 'label': 'NP-Spatial'},
  {'text': 'Cobourg', 'start': 22, 'end': 22, 'label': 'NP-Spatial'},
  {'text': 'Smalcalde', 'start': 24, 'end': 24, 'label': 'NP-Spatial'},
  {'text': 'Saxe-Gotha', 'start': 35, 'end': 37, 'label': 'NP-Spatial'},
  {'text': 'Long.', 'start': 39, 'end': 40, 'label': 'NP-Misc'},
  {'text': '28.15', 'start': 41, 'end': 43, 'label': 'NP-Misc'},
  {'text': 'lat.', 'start': 45, 'end': 46, 'label': 'Latlong'},
  {'text': '50.35', 'start': 47, 'end': 49, 'label': 'Latlong'},
  {'text': 'D.J.', 'start': 52, 'end': 53, 'label': 'Latlong'}]}
```

Prediction

```
[{'text': 'HILPERHAUSEN','start': 0,'end': 0,'label': 'Head'},
 {'text': 'Géog.','start': 3,'end': 4,'label': 'Domain-mark'},
 {'text': 'ville','start': 6,'end': 6,'label': 'NC-Spatial'},
 {'text': 'Allemagne','start': 8,'end': 8,'label': 'NP-Spatial'},
 {'text': 'en','start': 9,'end': 9,'label': 'Relation'},
 {'text': 'Franconie','start': 10,'end': 10,'label': 'NP-Spatial'},
 {'text': 'sur','start': 12,'end': 12,'label': 'Relation'},
 {'text': 'la Werra','start': 13,'end': 14,'label': 'NP-Spatial'},
 {'text': 'au','start': 16,'end': 16,'label': 'Relation'},
 {'text': 'comté','start': 17,'end': 17,'label': 'NC-Spatial'},
 {'text': 'Henneberg','start': 19,'end': 19,'label': 'NP-Spatial'},
 {'text': 'entre','start': 21,'end': 21,'label': 'Relation'},
 {'text': 'Cobourg','start': 22,'end': 22,'label': 'NP-Spatial'},
 {'text': 'Smalcalde','start': 24,'end': 24,'label': 'NP-Spatial'},
 {'text': 'Saxe-Gotha','start': 35,'end': 37,'label': 'NP-Spatial'},
 {'text': 'Long. 28. 15. lat. 50. 35','start': 39,'end': 49,'label': 'Latlong'}]
```

Ground truth

Entities not found

Wrong classification

Wrong boundaries

Conclusion, Reflection and perspectives

- Experiment other architectures for the named entities recognition
- Modify our GeoEDdA dataset to have a better balance between the different classes among sets (train, validation, and test)
- Hybrid Model (Base BERT+Grammatical rules)
- Clean code and release on Github

Questions ?

LABEX ASLAN
UNIVERSITÉ DE LYON