UkraiNER: A New Corpus and Annotation Scheme Towards Comprehensive Entity Recognition

Lauriane Aufrant & Lucie Chasseur

LREC-COLING 2024

Named Entity Recognition

Locating **names** of entities such as persons, locations, organizations... and **typing** them

Towards more comprehensiveness in NER

She invited their CEO to visit Princeton University, although students were off to see the NATO representative.

Towards more comprehensiveness in NER

She invited their CEO to visit Princeton University$_{LOC}$, although students were off to see the NATO$_{ORG}$ representative .

- **Flat NER** (CoNLL 2003): sequence labeling task, non-overlapping names

Towards more comprehensiveness in NER

She invited their CEO to visit Princeton$_{LOC}$ University$_{LOC}$, although students were off to see the NATO$_{ORG}$ representative .

- **Flat NER** (CoNLL 2003): sequence labeling task, non-overlapping names
- **Nested NER** (ACE 2004): all occurrences of names

Towards more comprehensiveness in NER

She invited their CEO to visit Princeton$_{LOC}$ University$_{LOC}$, although students were off to see the NATO$_{ORG}$ representative .

- **Flat NER** (CoNLL 2003): sequence labeling task, non-overlapping names
- **Nested NER** (ACE 2004): all occurrences of names
  Captures more information:
  ↪ "US Secretary of State": person & geopolitical entity
  ↪ "the Lufthansa Executive Board": the company & its board

Towards more comprehensiveness in NER

She$_{PER}$ invited their CEO$_{PER}$ to visit Princeton$_{LOC}$ University$_{LOC}$, although students were off to see the NATO$_{ORG}$ representative .

- **Flat NER** (CoNLL 2003): sequence labeling task, non-overlapping names
- **Nested NER** (ACE 2004): all occurrences of names
  Captures more information:
  ↪ "US Secretary of State": person & geopolitical entity
  ↪ "the Lufthansa Executive Board": the company & its board
- **Extended NER** (in part ACE, then Quaero 2011): pronominal or nominal mentions, more entity types

Towards more comprehensiveness in NER

She$_{PER}$ invited their CEO$_{PER}$ to visit
Princeton$_{LOC}$ University$_{LOC}$, although students were off
to see the NATO$_{ORG}$ representative .

- **Flat NER** (CoNLL 2003): sequence labeling task, non-overlapping names
- **Nested NER** (ACE 2004): all occurrences of names
  Captures more information:
  ↪ “US Secretary of State”: person & geopolitical entity
  ↪ “the Lufthansa Executive Board”: the company & its board
- **Extended NER** (in part ACE, then Quaero 2011): pronominal or nominal mentions, more entity types
  ↪ entities as the foundation for building knowledge bases

Towards more comprehensiveness in NER

She$_{PER}$ invited their CEO$_{PER}$ to visit Princeton$_{LOC}$ University$_{LOC}$, although students$_{GROUP}$ were off to see the NATO$_{ORG}$ representative$_{PER}$.

- **Flat NER** (CoNLL 2003): sequence labeling task, non-overlapping names
- **Nested NER** (ACE 2004): all occurrences of names
  Captures more information:
  ↪ "US Secretary of State": person & geopolitical entity
  ↪ "the Lufthansa Executive Board": the company & its board
- **Extended NER** (in part ACE, then Quaero 2011): pronominal or nominal mentions, more entity types
  ↪ entities as the foundation for building knowledge bases

Contributions

- A new annotation scheme: comprehensive entity recognition (but lightweight)

- Associated annotation tool, with automated suggestions

- A new corpus: UkraiNER, 10k French sentences in the geopolitical news domain

Annotation scheme for Comprehensive Entity Recognition

Main principles

Combining broad coverage, with lightweight annotations and a flexible format:

1. Broad set of (coarse) entity types
2. All entities in those classes, all mentions of those entities
3. Rich annotations to track unconventional mentions and enable filtering

CoNLL-UI format

```
# text = Biden met the European ambassadors last night.
# entity: 1 PERSON named complete main −− > Biden
# entity: 3-4-5 GROUP non-named complete main −− > the European ambassadors
# entity: 6-7 DATE non-named complete main −− > last night
1   Biden         Biden         PROPN   NNP   ···   ···
2   met           meet          VERB    VBD   ···   ···
3   the           the           DET     DT    ···   ···
4   European      european      ADJ     JJ    ···   ···
5   ambassadors   ambassador    NOUN    NNS   ···   ···
6   last          last          ADJ     JJ    ···   ···
7   night         night         NOUN    NN    ···   ···
8   .             .             PUNCT   .     ···   ···
```

Nested entities

[the President of [the lower house of [the Parliament $]_{O}$ $]_{O}$ $]_{P}$

[[the French President $]_{subd}$ [Jacques Chirac $]_{subd}$ $]_{main}$

[[France $]_{main}$'s President $]_{main}$, [Jacques Chirac $]_{main}$

[[the European Union $]_{subd}$ ([EU $]_{subd}$) $]_{main}$

Non-named entities

- Includes entities with no name (“*earlier*”) or name not mentioned (“*a former senator*”)
- Not only proper nouns: noun phrase, pronoun, pronominal phrase, adverb, numerals and symbols, free relative clauses…
- Excludes relational references to an entity (“*my*”, “*French*”)
- Mentions labeled as named or non-named

Discontinuous and incomplete entities

[the French]$_{\textit{incomplete}}$ and [British ambassadors]

[the French President] and [Prime Minister]$_{\textit{incomplete}}$

... said [M.]$_{\textit{incomplete}}$

Entity types

- **Person:** any natural or fictional individual (“*Pete’s wife*”, “*a tall man*”)
- **Location:** room, building, street, city, country...
- **Date:** dates, times, periods (absolute or relative)
- **Event:** public event, meeting, press conference, official speech... (including planned events)
- **Organization:** structured entities with legal status (companies, associations, administrations...) or without (collectives, bands...)
- **Group:** unstructured set of individuals (“*the victims*”) or organizations (“*the political parties*”), aggregates (“*civil society*”)
- **Product:** objects, equipment, software, websites, documents, laws...
- **Other:** distances, ages, percentages, currencies...

A first corpus for CER

UkraiNER

- News in French, from *Le Monde*
- Live feed about Ukraine on 19-28 February 2022 and 27-30 March 2022
- 1,555 news briefs, 10,604 sentences, 43,623 entities

UkraiNER

- News in French, from *Le Monde*
- Live feed about Ukraine on 19-28 February 2022 and 27-30 March 2022
- 1,555 news briefs, 10,604 sentences, 43,623 entities

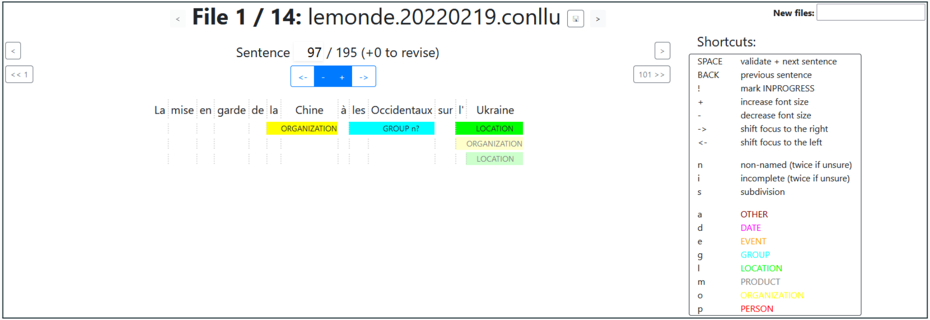
✓ More resources for French

✓ Manually annotated

✓ On a contemporary topic

Data preparation

- Scraping, pre-processing with UDPipe (tokenization, tagging, parsing)
- One main annotator + secondary annotator for quality control
- Iterative refinement of CER guidelines along the annotation process
- First annotations revised after completion
- In-house annotation tool with custom shortcuts and automated suggestions (based on history with exact text match)



File 1 / 14: lemonde.20220219.conllu

New files:

Sentence 97 / 195 (+0 to revise)

La mise en garde de la Chine à les Occidentaux sur l' Ukraine

ORGANIZATION GROUP n? LOCATION

ORGANIZATION

LOCATION

Shortcuts:

SPACE	validate + next sentence
BACK	previous sentence
!	mark INPROGRESS
+	increase font size
-	decrease font size
->	shift focus to the right
<-	shift focus to the left
n	non-named (twice if unsure)
i	incomplete (twice if unsure)
s	subdivision
a	OTHER
d	DATE
e	EVENT
g	GROUP
l	LOCATION
m	PRODUCT
o	ORGANIZATION
p	PERSON

Entity statistics

entity type	# entities	%	% named	% incomplete	length
all	43,623		44.9	1.3	2.6 ±1.9
PERSON	8,373	19.2	53.2	0.7	2.6 ±2.0
ORGANIZATION	10,869	24.9	71.7	0.2	2.6 ±1.8
GROUP	5,978	13.7	**0.3**	**4.8**	2.6 ±1.8
LOCATION	8,457	19.4	59.3	1.1	2.3 ±1.8
DATE	4,195	9.6	**39.9**	0.2	2.3 ±1.4
EVENT	2,366	5.4	**1.7**	0.5	**4.1** ±2.4
PRODUCT	3,119	7.2	**16.9**	2.0	2.7 ±1.7
OTHER	266	0.6	**22.2**	2.6	2.8 ±1.3

Baseline experiments

- Evaluation of **CNN-NER**, spaCy, Locate-and-Label
- F1 with exact match, micro and macro
- Comparing full (CER) and filtered (nested NER) annotations
- Train = Feb 22-28 + Mar 27-29, dev = Feb 19-21, test = Mar 30

Results (CNN-NER, micro)

train→test	Precision	Recall	F1
CER→CER	81.7	83.0	82.3
NNER→CER	88.8	42.4	57.4
NNER→NNER	89.1	85.8	87.4

Results (CNN-NER, macro, CER→CER)

	Precision	Recall	F1
PERSON	94.8	95.4	95.1
ORGANIZATION	87.3	86.3	86.8
GROUP	68.5	76.0	72.0
LOCATION	83.4	84.8	84.1
DATE	72.3	80.7	76.3
EVENT	57.4	50.4	53.6
PRODUCT	62.1	63.3	62.7
macro	75.1	76.7	75.8

Results (CNN-NER, macro, NNER→NNER)

	Precision	Recall	F1
PERSON	99.4	97.5	98.4
ORGANIZATION	90.0	85.0	87.4
LOCATION	76.7	81.6	79.0
DATE	96.4	81.4	88.2
PRODUCT	60.5	55.8	57.5
macro	84.6	80.2	82.1

Take-home messages

- NER guidelines cover a minor part of entity information

  ↪ More comprehensiveness needed for purposes like knowledge base extraction

- New annotation scheme: Comprehensive Entity Recognition

  ↪ Lightweight guidelines, flexible format, associated annotation tool

- A first corpus with that scheme: UkraiNER

  ↪ 10k French sentences in the geopolitical news domain

Thank you for watching!

For any question, comment, suggestion: first.last@inria.fr