Structured Named Entities in two distinct press corpora: Contemporary Broadcast News and Old Newspapers

Sophie Rosset[α] **Cyril Grouin**[α] Karën Fort[β,γ]
Olivier Galibert[δ] Juliette Kahn[δ] Pierre Zweigenbaum[α]

[α]LIMSI–CNRS, France [β]INIST–CNRS, France [γ]LIPN, France [δ]LNE, France

LAW 2012: 6th Linguistic Annotation Workshop
제주도, 대한민국 *(Jeju-do, Korea)* – July 12th, 2012

Introduction

Context

- **Quaero Project:**
  - Extracting information from news:
    - Proposal of a definition for **extended** and **structured** named entities; guidelines → (Rosset et al. 2011);
    - Annotation of two press corpora (1.5 million of words each one) used in two evaluation campaigns.
- **Corpus annotation:**
  - 2011: Broadcast News (BN) corpus, radio and television shows (Grouin et al. 2011; Galibert et al. 2011);
  - 2012: Old Press (OP) corpus, French newspapers from December 1890 (Galibert et al. 2012).
- Aims of this work: to compare annotations in both corpora.

Introduction

Named Entities

Text element classifiable on a semantic level:

- MUC-6: *person, location, organization*
- Numerical types: *date, time, money*
- Existing proposals:
  - finer-grained classes (*person* → *politician, location* → *city*);
  - new class: *product*, hierarchy w/ 200 types (Sekine 2004);
  - to fit historical data: *ships, regiments, railroads* (American Civil War).

Original objective

Answer to basic questions: Who? What? Where? When?

Extended Named Entities

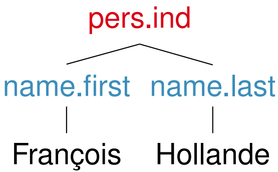
Our definition

- New types (*products*, *functions*),
- New coverage (expressions w/o proper nouns allowed),
- Structuring of the entities:
  - **Hierarchy:** types/subtypes taxonomy;
    - Type *person*:
      — Subtype *individual*: pers.ind
      — Subtype *collective*: pers.coll
    - Special subtypes:
      — *.oth (other subtype than those proposed)
      — *.unk (I don't know wich subtype to use).
  - **Compositionality:** entity composed of
    - types/subtypes (out of 31),
    - components (out of 30).

Extended Named Entities

Compositionality

Each entity type includes at least one component:



pers.ind

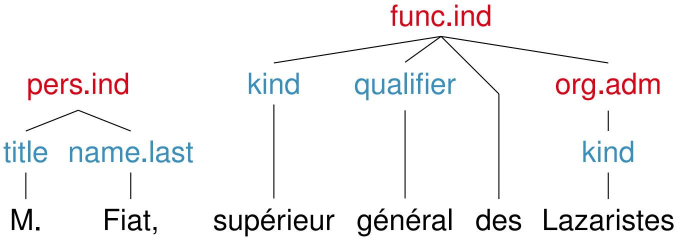
name.first name.last

François Hollande

Extended Named Entities

Compositionality

Another entity can act as a component:



pers.ind

title name.last

M. Fiat,

func.ind

kind qualifier org.adm

kind

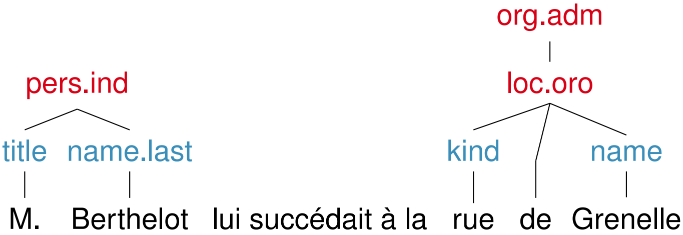
supérieur général des Lazaristes

Mr Fiat, General Superior of the Lazarists

Extended Named Entities

Metonymy and Antonomasia

An entity type can be used to refer to another type:



org.adm

pers.ind

loc.oro

title name.last

kind name

M. Berthelot lui succédait à la rue de Grenelle

Mr Berthelot was succeeding him at rue de Grenelle

“rue de Grenelle” = “Ministry of Education”

Some numbers

	BN corpus		OP corpus	
	Training	Test	Training	Test
# show/pages	188	18	231	64
# words	1,291,225	108,010	1,297,742	363,455
# entity types mentions	113,885	5,523	114,599	33,083
# entities w/ correction	—	—	4,258	1,364
# components mentions	146,405	8,902	136,113	40,432
# components w/ correction	—	—	71	22

Adaptation of the annotation

From Broadcast news to Old press

- OCRed Old Press corpus characteristics:
  - some remaining incorrectly recognized characters;
  - fixed-size columns from the original formatted text:
    → some remaining line breaks and hyphenations.
- Annotation adaptation to the Old Press corpus:
  - **Attribute “correction”**
    → annotators corrected incorrectly recognized entities:
    <loc.adm.town correction="d’Alger"> d’Algor </loc.adm.town>
  - **Component “noisy-entities”**
    → one or several entities combined due to a segmentation error (involves an entity boundary):
    <noisy-entities correction=”M. Montmerqué, ingénieur”>
    M. Montmerqué,ingénieur
    </noisy-entities>

Annotation evaluation

Creation of a mini reference corpus

- **Selection** of a sub-corpus from the training corpus
- **Annotation** by 2 teams of 2 annotators (A_1, A_2, B_1, B_2)
- **Adjudication:**
  1. within each team: A_1+A_2 / B_1+B_2
  2. from the previous ones: A+B
  3. with the annotated sub-corpus: AB+sub-corpus
     → **mini-reference corpus**.

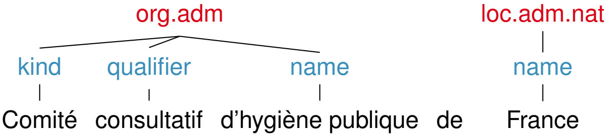
Inter-Annotator Agreement

Which markables? See (Grouin et al. 2011)	BN	OP
F-measure *(highest possible bound)*	0.845	0.799
All annotated entities as markables (κ, *lowest possible bound)*	0.713	0.647

Comparisons

Broadcast News vs. Old Press annotation campaigns

- **Source material** (more problems in OP corpus):
  - OCR errors that do not appear:
    → "*touché*" (touched) instead of "*Fouché*" (last name)
  - combined entities: "*M. Montmerqué,ingénieur*"
- **Language** (OP corpus is more difficult):
  - Specific languages: religious language, abbreviations;
  - Cultural context: geographical divisions from 1890.
    → *Tonkin:* country (loc.adm.nat) or region (loc.adm.reg)?
  - Annotation difficulties: boundary delimitation more difficult:



org.adm loc.adm.nat

kind qualifier name name

Comité consultatif d'hygiène publique de France

Consultative committee for public hygiene of France

Comparisons

Broadcast News vs. Old Press corpora

- **Statistical test** (Welch Two Sample t-test) to compare distribution of types across the corpora:

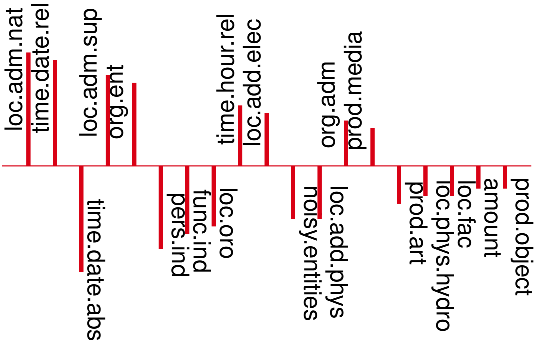


Figure: 19 entity types with $p < 0.001$, ranked by decreasing order of significance (top: BN corpus; bottom: OP corpus)

Comparisons

Broadcast News vs. Old Press corpora

- **Statistical test** (Welch Two Sample t-test) to compare distribution of components across the corpora:

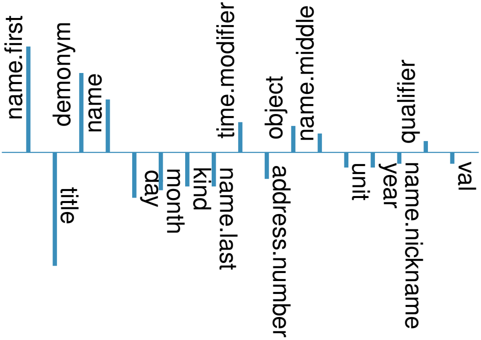


Figure: 17 components with $p < 0.001$, ranked by decreasing order of significance (top: BN corpus;bottom: OP corpus)

Comparisons

Structure differences across corpora

PATTERN	BN	OP
Type <pers.*> (person)		
- composed of <name.first/> and <name.last/>	52%	6%
- includes a <name.first/>	69%	19%
- composed of <title/> and <name.last/>	2%	34%
- includes a <title/>	8%	44%
Type <org.*> (organization)		
- <org.adm> <kind/>	6%	29%

Comparisons

Broadcast News vs. Old Press corpora

- **Automatic classification** based upon the distribution of types and components (73 tag ratios) across the corpora:

	FP	FN	FP+FN	Accuracy
One Rule	22	12	34	0.919
Decision Tree	2	5	7	0.983
Naïve Bayes	2	1	3	0.993
SVM	0	0	0	1.000

Table: Classification based on tag ratio (Weka toolbox)

- The ratios are discriminant enough to determine the corpus a document belongs to.

Conclusion and perspectives

- Same annotation scheme used in two corpora:
  - similar overall sizes (# tokens, # types and components)
  - but different annotation times.
- Comparisons made possible due to the structured definition;
- Human annotation process more difficult in OP;
- Future work:
  - further studies of comparison,
  - detecting relations between information,
  - new corpora annotation (w/ parallel FRE/ENG corpora).
- The corpora will soon be made available for free to the scientific community through ELDA catalogue.

Acknowledgements

This work was partly realized as part of:

- Quaero project funded by OSEO, French State agency for innovation
- ETAPE project funded by ANR, French National Agency for Research

- Galibert O, Rosset S, Grouin C, Zweigenbaum P, and Quintard L.
  Structured and extended named entity evaluation in automatic speech transcriptions. In Proc. of IJCNLP, Chiang Mai, Thailand. 2011.
- Galibert O, Rosset S, Grouin C, Zweigenbaum P, and Quintard L.
  Extended named entities annotation in ocred documents: From corpus constitution to evaluation campaign. In Proc. of LREC, Istanbul, Turkey. ELRA. 2012.
- Grouin C, Rosset S, Zweigenbaum P, Fort K, Galibert O, and Quintard L.
  Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In Proc. of the Fifth Linguistic Annotation Workshop (LAW-V), Portland, OR. Association for Computational Linguistics. 2011.
- Rosset S, Grouin C, and Zweigenbaum P.
  Entités Nommées Structurées : guide d'annotation Quaero.
  LIMSI–CNRS, Orsay, France. 2011
  http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf.
- Sekine S.
  Definition, dictionaries and tagger of extended named entity hierarchy. In Proc. of LREC. ELRA. 2004.