

Extended and Structured Entities in the Quaero Program

Sophie Rosset Cyril Grouin Pierre Zweigenbaum

LIMSI-CNRS, France



ETAPE
Paris, February 6th, 2012

- 1 Named Entities Definition
- 2 Extension of Named Entities
- 3 Recap
- 4 Guidelines Production
- 5 Annotations
- 6 Conclusions et Perspectives
- 7 Évaluation

Named Entities

Named Entities : pieces of text one can semantically classify.

Objective

Answer to some basic questions: *Who? What? Where? When? How? Why?*

Different Types (cf. MUC6)

- person names (*Zina Aïtlamin Nesrin El-Hamdaoui*) ;
- location names (*Brest*) ;
- organization names (*ONU*) ;
- amount and other quantities (*900 millions de dirhams*) ;
- temporal expressions like dates and duration.

References

- ESTER2 (Ester2, 2009) and its working group;
- PhD of Maud Ehrmann (Ehrmann, 2008) ;
- Temporal expression definition from TIMEX3 (Timex, 2009) ;
- PhD of Michaël Tran (Proper Nouns) (Tran, 2006) ;
- Named Entities Hierarchy from S. Sekine (Sekine, 2004a; Nadeau & Sekine, 2007).
- and many others

- 1 Named Entities Definition
- 2 Extension of Named Entities**
- 3 Recap
- 4 Guidelines Production
- 5 Annotations
- 6 Conclusions et Perspectives
- 7 Évaluation

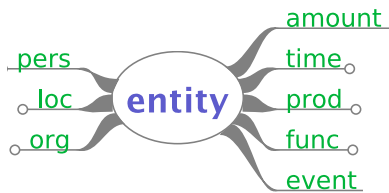
Objectives

- Knowledge database constitution:
 - Information extraction in *news* data;
 - Entities relation identification.
- ⇒ These ENE are the linchpin of the knowledge database.

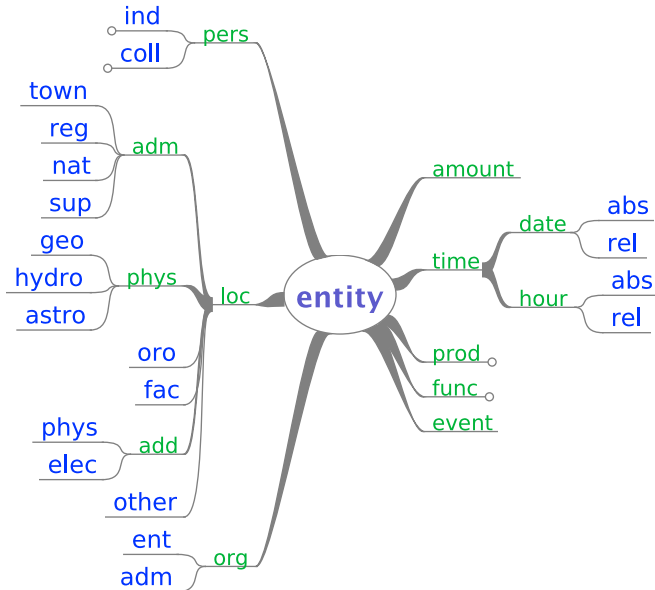
Definition extension

- Extension to new types (*products* (Sekine, 2004b), *functions* (Galliano *et al.*, 2009)) and to expressions without any proper noun;
 - Structuring of the entities:
 - **hierarchy** (types and sub-types) and
 - **composition** (specific and transverse components).
- ⇒ An entity has a type (and sub-type) and each elements of this entity is specified (component).

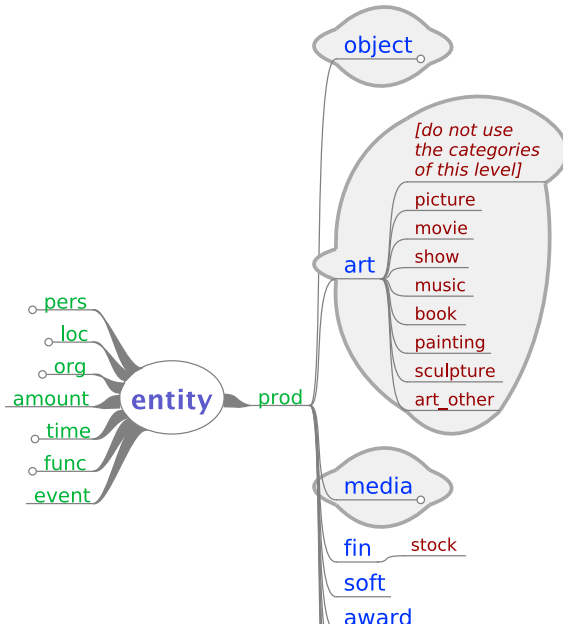
Overall presentation: Classical entities and additional entities



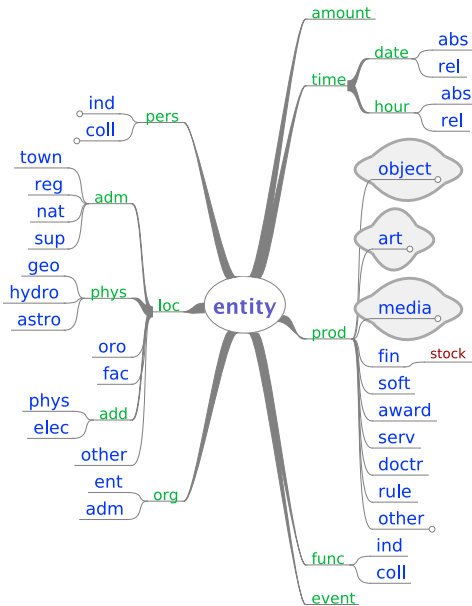
Overall presentation: Classical entities



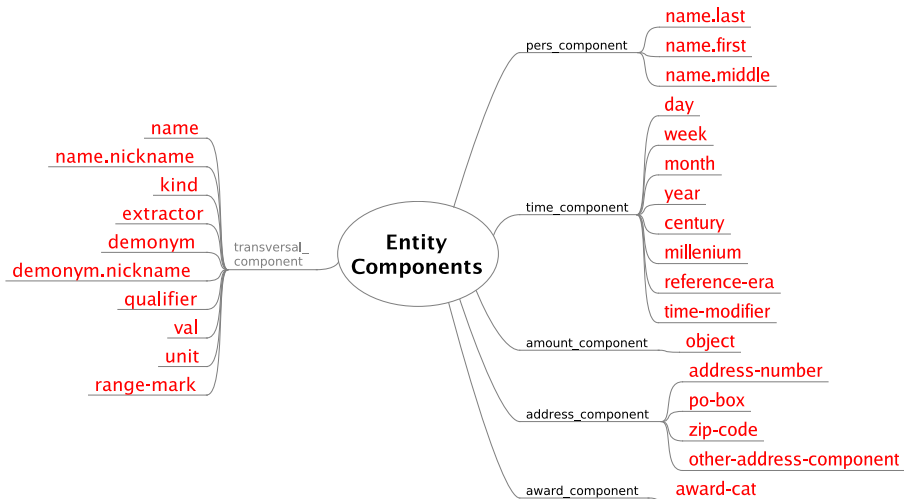
Overall presentation: Additional antities



Overall presentation: Hierarchy

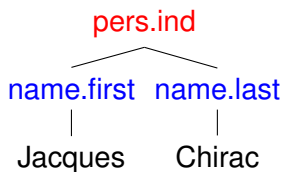


Overall presentation: Composition (generality)



Composition

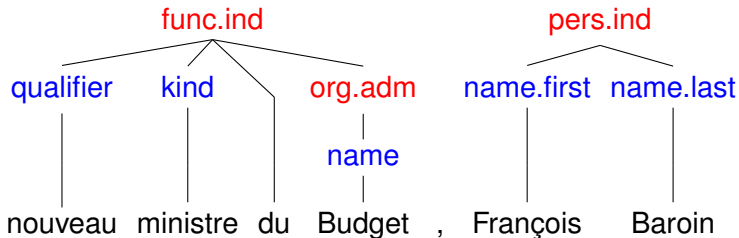
❶ A sub-type contains a component:



⇒ Each **sub-type** contains at least one **component**.

Composition

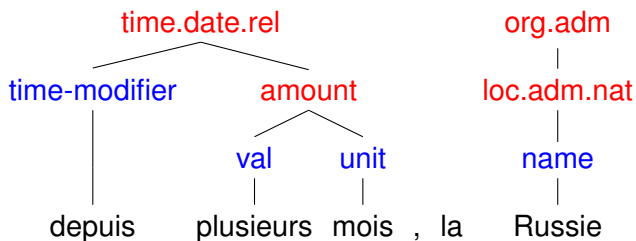
② A type includes another type: a sub-type can be used as a component.



⇒ the type **function** (**func.ind**) integrates a type **organisation** (**org.adm**).

Composition

③ Metonymy and antonomasia: an entity type is used to refer to another entity type:

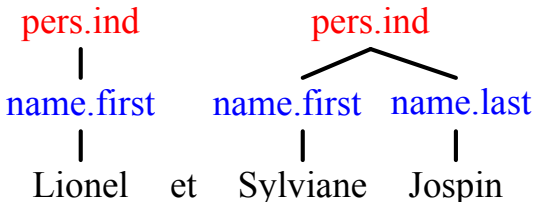


Boundaries

The scope of entities excludes:

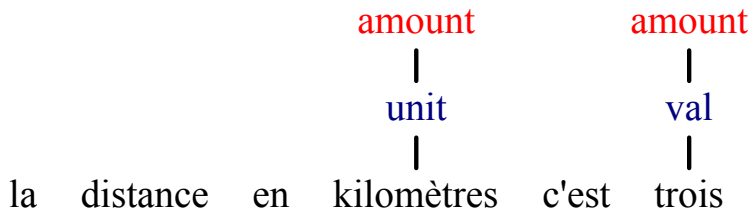
- relative clauses,
- subordinate clauses,
- interpolated clauses

The annotation of an entity must end before these clauses.



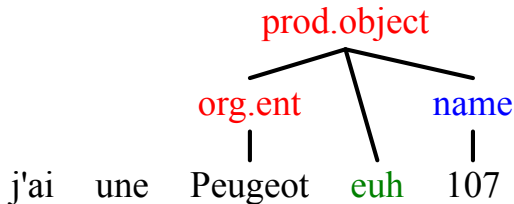
Dislocation

When a dislocation occurs, each part of the entity is annotated separately



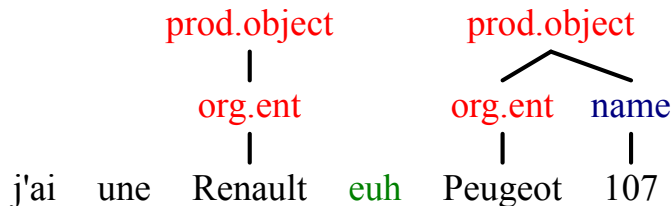
Disfluences

If an hesitation does not correct the entity, then it is part of the entity:



Disfluences

If an hesitation corrects the entity, then there is two entities



- 1 Named Entities Definition
- 2 Extension of Named Entities
- 3 Recap**
- 4 Guidelines Production
- 5 Annotations
- 6 Conclusions et Perspectives
- 7 Évaluation

Hierarchy (25) + unk + other

- **Person (2):** individual person, group of persons;
- **Location (5):** administrative location, physical location, facilities, oronyms, address;
- **Organization 2):** administration, service;
- **Time (4):** absolute and relative date, absolute and relative hour;
- **Amount;**
- **Product (9):** manufactured object, transportation route, financial products, doctrine, law, software, art, media, award;
- **Function (2):** individual function, collectivity of functions;

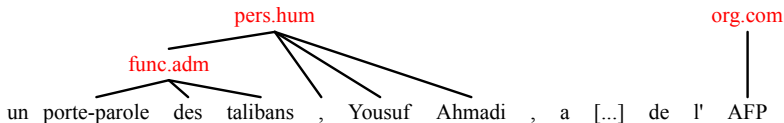
Composition (33)

- **Transverse components (7):** name of the entity, hypernym of the entity, qualifying adjective, demonym, number, unit, element in a series, range-mark, time modifier;
- **Specific components (26):**
 - *Individual Person:* first/middle/last name, title, pseudonym, nickname;
 - *Physical Address:* address number, PO box, zip code, other components;
 - *Date:* week, day, month, year, century, millenium, era;
 - *Amount:* object;
 - *Award:* award category.

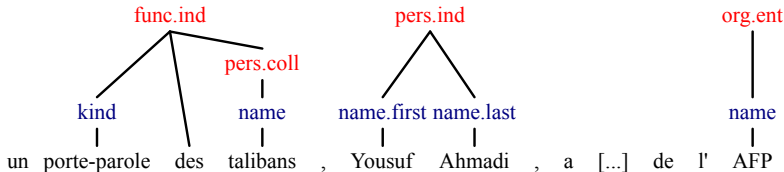
Ester2 vs Quaero

un porte-parole des talibans , Yousuf Ahmadi , a revendiqué l' attaque auprès de l' AFP

Ester 2



Quaero



- 1 Named Entities Definition
- 2 Extension of Named Entities
- 3 Recap
- 4 Guidelines Production**
- 5 Annotations
- 6 Conclusions et Perspectives
- 7 Évaluation

Process

3 researchers read all existing definition and associated remarks

New proposals: objectives definition, what annotated, what is realistic given the available money and time etc.

Double annotation by 2 researchers of different samples using the current extension

Corpus used during this step spoken (BN essentially) and written (news papers) data Quaero.

Spent time: 4 months, 3 researchers, at least 2 5-hours meeting a week

Important points

- Annotation is costly
- We defined some *optional annotation* which are managed by a normalization process
 - Evaluation is done on the normalized files
 - You should use normalized file for training
 - Non-normalized files may be richer than normalized ones (and then useful for future or other work)

- 1 Named Entities Definition
- 2 Extension of Named Entities
- 3 Recap
- 4 Guidelines Production
- 5 Annotations**
- 6 Conclusions et Perspectives
- 7 Évaluation

Process

Corpus : 188 shows for training and 18 for test (in Quaero !)
(broadcast news et broadcast conversations) ;

Double annotation by 4 annotators (linguist students, managed by ELDA) (10 man-month) ;

Regular evaluation during the annotation campaign, feedback analysis, guidelines correction (not the definition): discrepancies, residual errors corrected and more examples or rules added.

Annotated corpus used for the Quaero 2011 NE evaluation campaign and ... for ETAPE now

Some numbers

Inf. \ Data	Training	Test
# shows	188	18
# lines	43,289	5,637
# words	1,291,225	108,010
# types of entities	113,885	5,523
# distinct types	41	32
# components	146,405	8,902
# distinct components	29	22

Mini reference

- Mini reference corpus (400 lines randomly selected) in order to assess the quality of the overall corpus;
- Annotation done by 4 researchers (2 LIMSI et 2 INIST)
- Computation of κ coefficients given different *markables* (the random baseline is hard to evaluate with this kind of complex annotation) (Grouin *et al.*, 2011) :
 - Annotations LIMSI/INIST : $0.82 < \kappa < 0.91$;
 - Overall corpus vs. mini reference : $0.71 < \kappa < 0.85$.

But more important

- The overall annotations has been done on complete show (and not on lines...).
- The overall annotations are consistent

Conclusions et perspectives

- Modification of structure:
 - func.* can be transformed into components;
 - doable with the normalization
 - Merge of org.adm and org.ent ?
 - Some problems with the distinction between pers.coll and org.ent what should we do?
- Coming soon
 - Event annotation (PhD of Béatrice Arnulphy): Guidelines almost OK, should begin soon.
 - Relations between entities: we are working on them...

Références I

- 📄 EHRMANN M. (2008).
Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation.
PhD thesis, Université Paris 7 - Denis Diderot.
- 📄 ESTER2.
Annotation Entités Nommées, dates, heures et montants.
- 📄 GALLIANO S., GRAVIER G. & CHAUBARD L. (2009).
The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts.
In Proc of Interspeech 2009.
- 📄 GROUIN C., ROSSET S., ZWEIGENBAUM P., FORT K., GALIBERT O. & QUINTARD L. (2011).
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.
- 📄 NADEAU D. & SEKINE S. (2007).
A survey of named entity recognition and classification.
Linguisticæ Investigationes, **30(1)**.

Références II

- 📄 [SEKINE S. \(2004a\).](#)
Definition, dictionaries and tagger of extended named entity hierarchy.
In *LREC'04*, Lisbon, Portugal.
- 📄 [SEKINE S. \(2004b\).](#)
Definition, dictionaries and tagger of extended named entity hierarchy.
In *LREC'04*, Lisbon, Portugal.
- 📄 [TimeML Working Group.](#)
Guidelines for Temporal Expression Annotation for English for TempEval 2010.
- 📄 [TRAN M. \(2006\).](#)
Prolexbase. Un dictionnaire relationnel multilingue de noms propres : conception implantation et gestion en ligne.
PhD thesis, Université Francois Rabelais de Tours.

- 1 Named Entities Definition
- 2 Extension of Named Entities
- 3 Recap
- 4 Guidelines Production
- 5 Annotations
- 6 Conclusions et Perspectives
- 7 Évaluation**

Mesures d'évaluation classiques – 1

- **Rappel** (mesure de quantité) :

$$\text{Rappel} = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux négatifs}}$$

- **Précision** (mesure de qualité) :

$$\text{Précision} = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux positifs}}$$

- **F-mesure** (moyenne pondérée du Rappel et de la Précision) :

$$\text{F-mesure} = \frac{(1 + \beta^2) \times \text{Précision} \times \text{Rappel}}{\beta^2 \times \text{Précision} + \text{Rappel}}$$

Mesures d'évaluation classiques — 2

Le **Slot Error Rate** (Makhoul, 1999) complète ces mesures et prend en compte :

- **Suppression (D)** : entités dans Ref mais pas dans Hyp ;
- **Insertion (I)** : entités dans Hyp mais pas dans Ref ;
- **Types et frontières (TF)** : entités de Hyp avec erreur de type ET de frontière ;
- **Types (T)** : entités de Hyp avec erreur de type ;
- **Frontières (F)** : entités de Hyp avec erreur de frontière ;
- **Entités attendues (R)** : entités présentes dans Ref.

$$\text{Slot Error Rate} = \frac{D + I + TF + 0,5 \times (T + F)}{R}$$

Protocole d'évaluation

- Corpus de transcriptions de journaux et d'émissions ;
- Annotations de **référence** sur la **transcription manuelle** ;
- Annotations des **systèmes des participants** sur :
 - la **transcription manuelle**
 - **4 transcriptions automatiques** :
 - 1 sorties de 3 systèmes différents (ASR1, ASR2, ASR3) ;
 - 2 sortie enrichie du premier système (ASR1+) : ponctuation et majuscule en début de phrase.

⇒ Besoin d'aligner référence et annotations des participants.

Alignement

- Appariement entre les entités de Hyp et de Ref ;
- Définition d'un coût d'appariement (SER) ;
- Association entre les entités de Hyp et de Ref au plus faible coût ;
- Algorithme : Viterbi (Programmation dynamique).

⇒ Cadre générique flexible où le coût dirige l'évaluation.

Projection sur une transcription automatique — 1

- Problème : on n'a pas d'annotation des entités nommées étendues sur une sortie ASR ;
- Question : *voulons-nous annoter les mots reconnus ou les mots effectivement prononcés ?*
- *Ce qui a été dit* a davantage de sens d'un point de vue applicatif.

⇒ besoin de projeter la référence manuelle sur une sortie ASR.

Projection sur une transcription automatique — 2

- Construire un alignement temporel de la transcription manuelle ;
- Extraire à partir de cet alignement le temps pour les annotations de référence ;
- Sélectionner une taille d'intervalle (médiane des longueurs des mots reconnus).

⇒ Nous avons l'information temporelle, nous devons retourner aux mots reconnus.

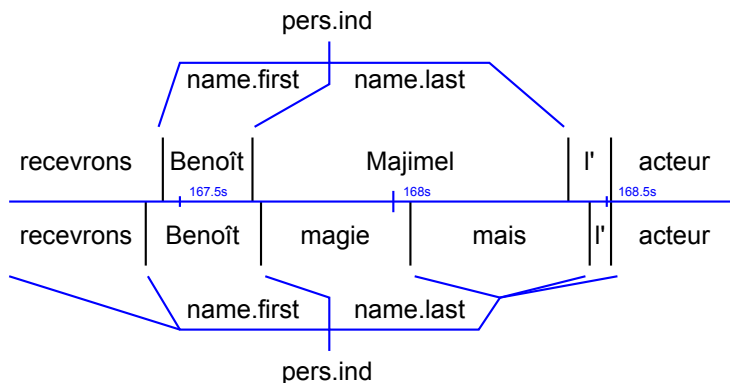
- Trouver les frontières des mots reconnus entrant dans les intervalles (début et fin).

⇒ Il y a souvent de multiples frontières.

- Construire une référence *floue* où des frontières multiples sont possibles.

Projection sur une transcription automatique — 3

Exemple d'une référence floue obtenue par alignement temporel de la référence propre (haut) sur ASR (bas) :



Résultats

- 3 participants (P1, P2, P3)
- Données d'entraînement = Ester1 + Ester2
- Données de test : DEV+Test évaluation Quaero ASR 2010

Perte absolue en terme de SER.

	Manuel	ASR1	ASR1+	ASR2	ASR3
Broadcast News					
WER		16.32%	16.32%	18.77%	24.06%
P1	42.7%	-12.5%	-10.0%	-15.8%	-18.7%
P2	29.7%	-18.8%	-24.1%	-22.5%	-23.8%
P3	39.1%	-16.5%	-15.4%	-21.2%	-22.7%
Broadcast Conversations					
WER		23.34%	23.34%	22.99%	29.18%
P1	55.3%	-32.6%	-34.6%	-23.0%	-33.9%
P2	37.0%	-36.9%	-42.0%	-29.6%	-36.0%
P3	43.0%	-46.3%	-40.3%	-38.2%	-41.1%

Résultats

- Sous-types plutôt bien annotés (F_1 -mesure) :

Sous-type	P1	P2	P3	Référence
pers.ind	0.834	0.893	0.900	0.991
loc.adm.town	0.759	0.745	0.797	0.947

- Sous-types problématiques (F_1 -mesure) :

Sous-type	P1	P2	P3	Référence
func.ind	0.461	0.552	0.599	0.966
org.adm	0.418	0.448	0.360	0.911
org.ent	0.313	0.521	0.509	0.864
pers.coll	0.446	0.557	0.478	0.788