Université François-Rabelais Tours

LI
Laboratoire d'Informatique
EA 2101

CasEN : Une cascade de graphe pour la reconnaissance des entités nommée en français

Version réalisée dans le cadre
du projet Ortolang

Denis MAUREL, Nathalie FRIBURGER,
Université François-Rabelais de Tours

Quatre exemples d'annotation des entités nommées

MUC

Ester 2

Etape

Ortolang

Muc 7 (1997)

- En 1997, Chinchor définit les *Entités nommées* par 7 catégories divisées en 3 classes:
  - Enamex
    - personnes, lieux et organisations
  - Timex
    - dates et heures
  - Numex
    - pourcentages et valeurs monétaires

Exemple

- Migaud : "Il faut trouver de l'ordre de 33 milliards d'euros pour 2013". Premier président de la Cour des comptes et ancien député PS, Didier Migaud a remis...

Extrait du site du journal Le Monde, consulté le 2 juillet 2012

Exemple

- <enamex type="person">Migaud</enamex> : "Il faut trouver de l'ordre de <numex type="money">33 milliards d'euros</numex> pour <timex type="date">2013</timex>". Premier président de la <enamex type="organization">Cour des comptes</enamex> et ancien député <enamex type="organization">PS</enamex>, <enamex type="person">Didier Migaud</enamex> a remis...

Ester 2 (2009)

- Dans le cadre d'Ester 2, on utilise 7 catégories (divisées en 26 classes et 14 sous-classes):
  - personne
  - fonction
  - organisation
  - lieu
  - production humaine
  - date et heure
  - montant

Exemple

- Migaud : "Il faut trouver de l'ordre de 33 milliards d'euros pour 2013". Premier président de la Cour des comptes et ancien député PS, Didier Migaud a remis...

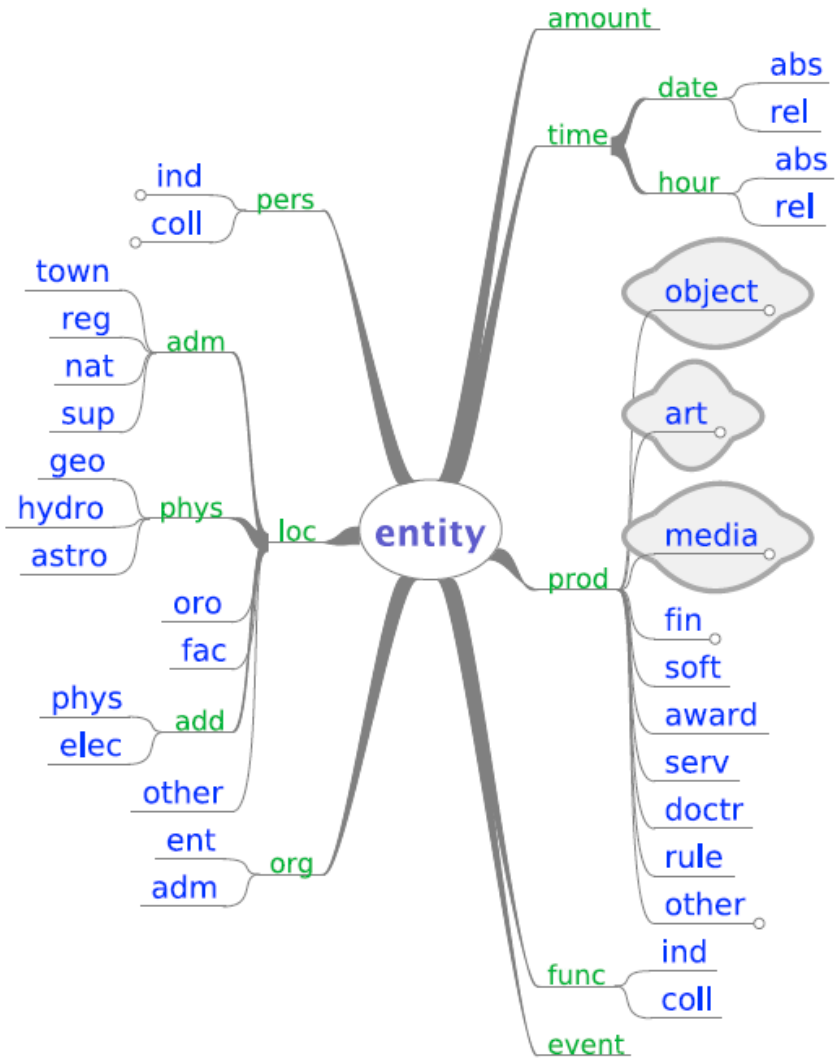
Extrait du site du journal Le Monde, consulté le 2 juillet 2012

Exemple

- <entity type="pers.hum">Migaud</entity> : "Il faut trouver de l'ordre de <entity type="amount.cur">33 milliards d'euros</entity> pour <entity type="time.date.abs">2013</entity>". <entity type="fonc.admi">Premier président de la <entity type="org.pol">Cour des comptes</entity></entity> et <entity type="fonc.pol">ancien député <entity type="org.pol">PS</entity></entity>, <entity type="pers.hum">Didier Migaud</entity> a remis...

Etape (2013)

8 catégories
25 classes
13 sous-classes



Exemple

- Migaud : "Il faut trouver de l'ordre de 33 milliards d'euros pour 2013". Premier président de la Cour des comptes et ancien député PS, Didier Migaud a remis...

Extrait du site du journal Le Monde, consulté le 2 juillet 2012

Exemple

- <pers.ind>Migaud</pers.ind> : "Il faut trouver <amount>de l'ordre de 33 milliards d'euros</amount> <time.date.abs>pour 2013</time.date.abs>". <fonc.ind>Premier président de la <org.adm>Cour des comptes</org.adm></fonc.ind> et <fonc.ind>ancien député <org.adm>PS</org.adm></fonc.ind>, <pers.ind>Didier Migaud</pers.ind> a remis…

Une complexification

Campagne	Nombre d'étiquettes	Pages du guide
Muc	7	23
Ester	37	25
Etape	54 (34+20)	86

Ortolang (2014)

- Présentation
  - Un choix de huit balises, conformes à la TEI:
    - personnes: <persName>
    - lieux: <placeName> et <geogName>
    - organisations: <orgName>
    - adresses: <address>
    - mesures: <measure>
    - temps: <date> et <time>

Exemple

- Migaud : "Il faut trouver de l'ordre de 33 milliards d'euros pour 2013". Premier président de la Cour des comptes et ancien député PS, Didier Migaud a remis...

Extrait du site du journal Le Monde, consulté le 2 juillet 2012

Exemple

- <persName>Migaud</persName> : "Il faut trouver de l'ordre de <measure>33 milliards d'euros</measure> pour <date>2013</date>". Premier président de la Cour des comptes et ancien député PS, <persName>Didier Migaud</persName> a remis...

Quelques difficultés

- La prise en compte du contexte linguistique (métonymie)
  - la <loc.adm.nat>France</loc.adm.nat> connaît son printemps le plus chaud depuis un siècle
  - en 2008, la <org.adm>France</org.adm> a réhabilité la Syrie…
  - elle aurait pu être l' occasion d'envoyer un un signal à cette <pers.coll>France</pers.coll> qui travaille

Exemples extraits de la campagne Etape

  - <address><lb/> 3 rue Désirée, Paris XX<hi rend="E">e</hi>, France</address>

Exemple extrait de l'ouvrage E163 de la collection Frantext

Quelques difficultés

- La prise en compte du contexte linguistique (métonymie)
  - prochaine séance de questions au gouvernement **à l'assemblée nationale**
  - prochaine séance de questions au gouvernement à l' <org.adm>assemblée nationale</org.adm>
  - prochaine séance de questions au gouvernement à l'<loc.fac>assemblée nationale</loc.fac>

Exemple extrait de la campagne Etape

Quelques difficultés

- L'ambiguïté d'interprétation
  - Monsieur le le premier ministre, **dans quelques jours** nous serons amenés à discuter…
  - Monsieur le le premier ministre, <amount>dans quelques jours</amount> nous serons amenés à discuter…
  - Monsieur le le premier ministre, <time.date.rel>dans quelques jours</time.date.rel> nous serons amenés à discuter…

CasEN

Version 2014

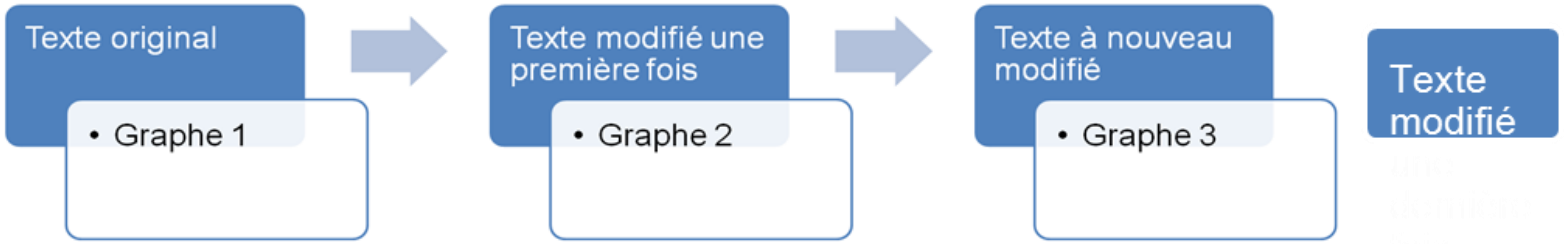
Téléchargeable à l'URL

http://tln.li.univ-tours.fr/Tln_CasEN.html

CasSys

- Un module pour la constitution et l'utilisation de cascades de graphes intégré à Unitex

  (Text/Apply CasSys Cascade…)



CasEN

- CasEN est une cascade de graphes dédiée à la reconnaissance des entités nommées en français
- 1ère place à la campagne d'évaluation Etape sur la tâche REN (textes transcrits manuellement)
- Une nouvelle version vient d'être développée avec le support du projet Ortolang

CasEN-Ortolang

- Présentation
  - Utilisation de trois dictionnaires
    - Dictionnaire prioritaire (ambiguïtés)
    - Dictionnaire de noms propres (Prolex-Unitex)
    - Dictionnaire de marqueurs-déclencheurs
  - Une cascade d'analyse de 106 graphes
  - Une cascade de synthèse de 31 graphes

CasEN-Ortolang

- Présentation
  - Ordre des graphes
    - Il est arrivé le 29 février de l'année 2008.

→par le graphe timeAnneesSiecle

    - Il est arrivé le 29 février de l'année 2008.

→par le graphe timeDateRelative

    - Il est arrivé le 29 février de l'année 2008.

→par le graphe timeDateAbsolue

    - Il est arrivé le 29 février de l'année 2008.

CasEN-Ortolang

- Présentation
  - Compléments entre graphes
    - rue du Général Leclerc
    - rue du 11 novembre 1918
    - Centre Georges Pompidou
    - hôpital Henri Mondor

CasEN-Ortolang

- Présentation
  - Plan approximatif de la cascade d'analyse
    - Outils (XML, particularité du corpus, phrases, nombres, formes figées…)
    - Mesures, dates, heures…
    - Personnes, fonctions, organisations
    - Lieux géographiques et administratifs
    - Personnes, fonctions, organisations, lieux
    - Adresses, bâtiments

CasEN-Ortolang

- Évaluation (*ouvrage E161 de la collection Frantext*)
  - #I = entités détectées par erreur
  - #D = entités totalement manquées
  - #T = typage incorrect
  - #E = balises mal placées
  - #S = entités détectées
  - #R = entités réelles

CasEN-Ortolang

- Évaluation (*ouvrage E161 de la collection Frantext* – 118 pages)
  - #I = entités détectées par erreur
  - #D = entités totalement manquées
  - #T = typage incorrect
  - #E = balises mal placées
  - #S = entités détectées
  - #R = entités réelles

#I	#D	#T	#E	#TE	#S	#R
191	437	89	68	48	2349	2595

CasEN-Ortolang

- Évaluation (*ouvrage E161 de la collection Frantext*)
  - SER (*slot error rate*): 29,1%

$$SER = \frac{\#I + \#D + 0{,}5 * \#T + 0{,}5 * \#E + \#TE}{\#R}$$

CasEN-Ortolang

- Évaluation (*ouvrage E161 de la collection Frantext*)
  - SER (*slot error rate*): 29,1%
  - Rappel: 83,2%

$$Rappel = \frac{\#S - \#I}{\#R}$$

CasEN-Ortolang

- Évaluation (*ouvrage E161 de la collection Frantext*)
  - SER (*slot error rate*): 29,1%
  - Rappel: 83,2%
  - Précision: 91,9%

$$Précision = \frac{\#S - \#I}{\#S}$$

CasEN-Ortolang

- Évaluation (*ouvrage E161 de la collection Frantext*)
  - SER (*slot error rate*): 29,1%
  - Rappel: 83,2%
  - Précision: 91,9%
  - Précision du typage: 86,0%

$$Précision\ du\ typage = \frac{\#S - (\#I + \#T + \#TE)}{\#S}$$

CasEN-Ortolang

- Évaluation (*ouvrage E161 de la collection Frantext*)
  - SER (*slot error rate*): 29,1%
  - Rappel: 83,2%
  - Précision: 91,9%
  - Précision du typage: 86,0%
  - Précision du balisage: 86,9%

$$Précision\ du\ balisage = \frac{\#S - (\#I + \#E + \#TE)}{\#S}$$

CasEN-Ortolang

- Exemples d'erreur
  - <s>La fratrie Jablonka est composée, dans<lb/> l'ordre, de Simje (né <date>en 1904</date>), <persName>Reizl</persName> (1907), Matès (1909),<lb/> Hershl (1915) et Henya (1917) - trois garçons et deux<lb/> filles, nés dans l'empire des tsars<hi rend="E">5</hi>.</s>

CasEN-Ortolang

- Exemples d'erreur
  - <s>La fratrie Jablonka est composée, dans<lb/> l'ordre, de Simje (né <date>en 1904</date>), <persName>Reizl</persName> (1907), Matès (1909),<lb/> Hershl (1915) et Henya (1917) - trois garçons et deux<lb/> filles, nés dans l'empire des tsars<hi rend="E">5</hi>.</s>

CasEN-Ortolang

- Corrections
  - Ajout de quatre prénoms dans le dictionnaire: Simje, Matès, Hershl et Henya
  - Modification de deux graphes: *persFamilleNom* et *timeDateNum*

CasEN-Ortolang

- Résultat corrigé
- <s>La fratrie <persName>Jablonka</persName> est composée, dans<lb/> l'ordre, de <persName>Simje</persName> (né en <date>1904</date>), <persName>Reizl</persName> (<date>1907</date>), <persName>Matès</persName> (<date>1909</date>),<lb/> <persName>Hershl (<date>1915</date>) et <persName>Henya</persName> (<date>1917</date>) - trois garçons et deux<lb/> filles, nés dans l'empire des tsars<hi rend="E">5</hi>.</s>

Fichiers résultats

Cascade d'analyse
{Chine,.entity+loc+admi}

- Sortie XML-CasSys

```
<csc>
<form>Chine</form>
<code>entity</code>
<code>loc</code>
<code>admi</code>
</csc>
```

Cascade de synthèse

- Sortie au format désiré

```
<placeName>
Chine
</placeName>
```

Merci !