Détection des entités nommées : un système symbolique open source à la disposition de la communauté

Nathalie Friburger, Denis Maurel

A l'origine

2000-2001 création de Cassys

. Système permettant de gérer une cascade de transducteurs (aujourd'hui intégré à Unitex)

. NER limitée aux personnes, lieux, organisations

2007-2010 participation de CasEN à Ester2, EPAC, Variling

. Extension aux entités : montants, événements, fonctions, productions humaines

Refonte de CasEN

- Architecture de la cascade
- Nommage des graphes
- Ecriture des graphes
- Outils de deboggage

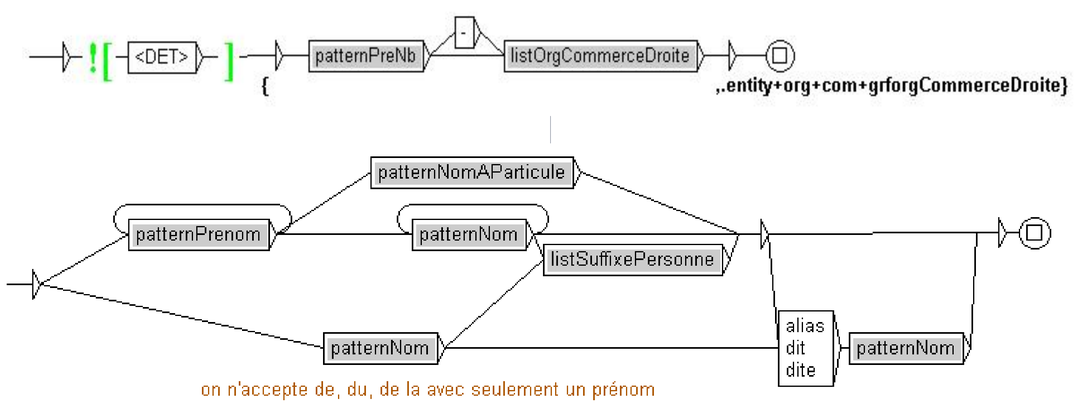
Pourquoi réécrire CasEN ?

.De nombreuses modifications apportées par des personnes différentes

.Maintenance difficile

.Nouvelles fonctionnalités d'Unitex : morphologie, contextes à exclure ou pas

.Modification du fonctionnement de CasSys / intégration à Unitex



Différents types de transducteurs

Fonction des transducteurs identifiable par le préfixe de leur nom

- **pers, loc, org, event, fonc, prod, amount, time** : reconnaissent des entités nommées correspondant au type indiqué (peuvent rassemble plusieurs transducteurs)
  ex : orgCommerceEtranger reconnait des organisations commerciales ayant un nom étranger
  ex : amount reconnait différentes mesures (monnaie, température, longueur, etc.).
- **tool** : transducteur outil.
  Ex : toolChercheSigleAvecPoints recherche les sigles contenant des points (*C.G.T.* au lieu de *CGT* aujourd'hui).
- **list** : transducteur de listes
- **pattern** : transducteurs masques, décrivent des motifs utilisant des codes Unitex pour sélectionner des séquences de mots (contraintes morphologiques, syntaxiques, lexicales etc.)
- **tag** : transducteur étiquette, permettent d'insérer des étiquettes dans des entités imbriquées

Les étiquettes « extérieures » , placées sur les ENs, ont la forme générale {xxxx,.y1+y2+…+y3}

Les traits y1, y2, …, yn peuvent être recherchés dans des transducteurs unitex par des masques plus ou moins spécifiques.

Ex : <y1>, <y2+y3>, <y2~y3>

{{Gerhard,.N+Prénom} {Aigner,.N+nom},.entity+pers+hum+grfpersPrenomNom}, le {secrétaire général,.entity+fonc+pol} de l'{UEFA,.entity+org+grforgMetier}, a dans ses placards des projets de {Ligue européenne,.entity+org+grforgDivers}.

. Ex : <pers+hum> → {{Gerhard,.N+Prénom} {Aigner,.N+nom},.entity+pers+hum+grfpersPrenomNom}
. <fonc> de <org> → { {secrétaire général ,.entity+fonc+pol} de l' {UEFA ,.entity+org+grforgMetier},.

Les transducteurs se complétent !

Ex : Les adresses postales
- Masques de personne (pour reconnaître rue du Général Leclerc)
- Masques de date (pour rue du 11 novembre 1918)

Les graphes des dates et de personnes placés avant le graphe des adresses.

Ex : organisations
- Centre Georges Pompidou ou l'hôpital Henri Mondor

Ces organisations seront donc reconnues après les graphes de personnes

Les transducteurs se concurrencent !

Principe de la cascade

- Utiliser dans les descriptions suivantes les motifs déjà détectés ou, au contraire, éviter un étiquetage non souhaité pour un motif déjà reconnu.
- L'ordre de passage de ces transducteurs est donc un paramètre important : ilôts de certitude en premier !

« Au pire de la crise, {à l'automne dernier,.entity+time+date+rel+grftimeDateRelative}, nous avons détenu jusqu'à 20 % de liquidités dans notre portefeuille », indique {{{ Denis,.N+Prénom} { Remacle,.N+nom},.entity+pers+hum}, {gérant d'{Amplitude Pacifique,.entity+org+com} ,.entity+job} ,.entity+pers+hum+grfpersPrenomNom}, une sicav de {La Poste,.entity+org+com+grforgDico}.{S}
…
« C'est à nos clients de décider s'ils souhaitent ou non consacrer une partie de leur patrimoine à l'{Asie,.entity+loc+admi+grflocPays} » , souligne {{Pierre,.N+Prénom} {Ciret,.N+nom},.entity+pers+hum+grfpersPrenomNom}, de la {Compagnie financière {{Edmond,.N+Prénom} {de Rothschild,.N+nom} ,.entity+pers+hum} ,.entity+org+com+grforgCommerceGauche}.
…
Ils ne peuvent pas, en revanche, faire l'impasse sur la {Bourse de {Hongkong , .entity+loc+admi},.entity+org+com+grforgCommerceGauche}, car cette place représente près de la moitié de la capitalisation boursière de la région.{S} Pour sa part, {{ Pierre-Alexis,.N+Prénom} { Dumont,.N+nom},.entity+pers+hum+grfpersPrenomNom}, de {State Street Banque,.entity+org+com+grforgCommerceDroite}, s'est réfugié sur le marché australien, relativement épargné par la tourmente.

{Théâtre Gérard-Philipe,.entity+org+div+grforgDivertissementSorties}, {59, {boulevard Jules-Guesde,.entity+loc+line}, 93000 {Saint-Denis,.entity+loc+ville},.entity+loc+addr+post+grflocAddr}.

Selon une étude de l'{Autorité de régulation des télécommunications,.entity+org+grforgDivers} ({ART,.entity+org+grforgOrgSuiviDeParentheses}), le taux d'équipement devrait dépasser les 50 % {en 2002,.entity+time+date+abs+grftimeAnneeSiecle}.

Prétraitements

- Découpage en phrase
- Dictionnaires propres à Casen + dictionnaires génériques d'unitex
- Pas de POS !

Pour faciliter le débogage de CasEN

Noms des graphes insérés (ex : *grftimeDateRelative)*

Résultats préliminaires

Journal Le Monde 12/08/1998
Taille du corpus : 18000 mots
Référence : 1656 entités nommées

Perspectives

- Annotation de différents corpus
- Evaluation de CasEN
- Création de 2 cascades :
  - L'une très précise
  - L'autre moins précise mais avec un meilleur rappel