RefGen : un système d'identification automatique des chaînes de référence

Laboratoire LiLPa (EA1339)
(Linguistique, Langues, Parole)
22 avenue René Descartes
67081 Strasbourg
http://lilpa.u-strasbg.fr/

RBS
Aéroparc 1,11 rue Icare
67836 Strasbourg Entzheim
Tanneries Cedex
www.rbs.fr

Laurence Longo, Amalia Todirascu
longo@unistra.fr, todiras@unistra.fr

1. Objectif

Développer un module de détection automatique des chaînes de référence

- visée applicative : intégration dans un outil de détection de thèmes pour optimiser un moteur de recherche

2. Contexte

Une chaîne de référence (CR) :

- marqueur linguistique qui indique une continuité ou une rupture thématique dans le discours (Cornish, 1995), (Goutsos, 1997)

- inclut au moins 3 expressions référentielles (ou maillons) référant la même entité du discours (Schnedecker, 2005)

Ex : « **Carla Bruni-Sarkozy** cache **son** petit ventre avec **son** châle à chaque fois qu'un photographe approche, **elle** ne dément pas quand on **lui** pose la question hors micro. » (Le Monde, 02/05/11)

Travaux existants en détection des CR :

- systèmes à base de règles (Mitkov, 2001)

- apprentissage à partir de corpus annotés (Ng et Cardie, 2002), (Hoste, 2005), (Denis et Baldridge, 2008)

RefGen applique une méthode de calcul de la référence basée sur :

- l'accessibilité d'(Ariel, 1990)
- la vérification de contraintes
- des paramètres dépendants du genre textuel (Longo et Todirascu, 2010)

3. Annotation automatique des expressions référentielles dans *RefGen*

- **Etiquetage du texte avec TTL (Ion, 2007) :**
  - o segmentation en *chunks* simples : groupes nominaux (Np), prépositionnels (Pp)
  - o informations morpho-syntaxiques fines (projet Multext (Ide et Véronis, 1994)) : temps, mode, personne, genre, nombre

- **Application de patrons symboliques** pour identifier des expressions plus informatives :
  - o les groupes nominaux complexes (CNp)
  - o les entités nommées (Ner) : organisations, personnes, fonctions, lieux

```
<w lemma="le" chunk="Np#1" ana="Da-fs">L'</w>
<w lemma="union" chunk="Np#1" ana="Ncfs" ner="NER#1, org">Union</w>
<w lemma="européen" chunk="Np#1, Ap#1" ana="Af-fs" ner="NER#1, org">européenne</w>
<w lemma="avoir" chunk="Vp#1" ana="Vaip3s">a</w>
<w lemma="adopter" chunk="Vp#1" ana="Vmps-s">adopté</w>
<w lemma="il" ana="Pp3ms" feat="imp">il</w>
<w lemma="y" ana="Pp3">y</w>
<w lemma="avoir" ana="Vaip3s">a</w>
<w lemma="peu" chunk="Ap#2" ana="R">peu</w>
<w lemma="de_le" chunk="CNp#5, Pp#1, Np#2" ana="Dg-mp">des</w>
<w lemma="acte" chunk="CNp#5, Pp#1, Np#2" ana="Ncmp">actes</w>
<w lemma="législatif" chunk="CNp#5, Pp#1, Np#2, Ap#3" ana="Af-mp">législatifs</w>
<w lemma="relatif" chunk="CNp#5, Pp#1, Np#2, Ap#3" ana="Af-mp">relatifs</w>
<w lemma="à+le" chunk="CNp#5, Pp#2, Np#3" ana="Dg-ms">au</w>
<w lemma="changement" chunk="CNp#5, Pp#2, Np#3" ana="Ncms">changement</w>
<w lemma="climatique" chunk="CNp#5, Pp#2, Np#3, Ap#4" ana="Af-ms">climatique</w>
```

Exemple de sortie enrichie dans *RefGen*

4. Algorithme d'identification des chaînes de référence dans *RefGen*

Plusieurs étapes pour identifier automatiquement les chaînes de référence :

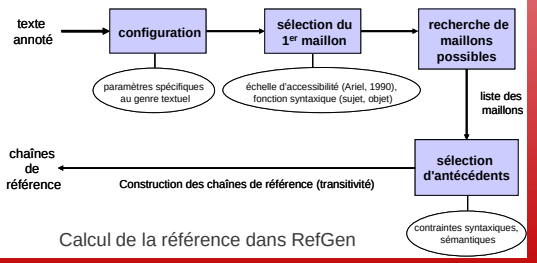
- **Sélection du premier maillon :** accessibilité et paramètres dépendants du genre textuel

ex : entités i, j dans [M. Pons]**i** rappelle que [J. Chirac]**j** [lui]**p** apparaît comme [[le candidat légitime]**k** de [son parti]**m**]**n**.

- **Sélection des autres maillons :**

a) paires antécédent-anaphore possibles :
ex : (i,p), (i, m), (i,k), (j,p), (j,m), (j,k), (j,n)

b) contraintes syntaxiques, sémantiques, lexicales à valider :
- o **fortes** (éliminatoires) : imbrication (m,n), arguments du même prédicat (j,p)
- o **faibles** (à valider au maximum) : correspondance genre/nombre (i,p)



Calcul de la référence dans RefGen

5. Evaluation

Corpus d'évaluation : rapports publics de la Commission Européenne, 7230 mots, sur le changement climatique

- Comparaison des annotations et du calcul de la référence entre les sorties automatiques de *RefGen* et une annotation manuelle

- Triple évaluation du corpus : avec les paramètres spécifiques au genre étudié (rapports publics), avec les paramètres du genre journalistique, sans aucun paramètre

	Annotations			CalcRef	
	Ner	CNp	II $_{imp}$	paires	chaînes de référence
rappel	0,85	0,87	0,91	0,69	0,58
précision	0,91	0,91	1	0,78	0,70
f-mesure	0,88	0,89	0,95	0,73	0,63
f-mesure (genre : journaux)				0,70	0,54
f-mesure (sans paramètre)				0,71	0,51

6. Perspectives

- Présentation d'un système d'identification automatique des CR : *RefGen*

- Architecture modulaire : ajout de règles symboliques possible pour identifier d'autres cas (anaphores plurielles , …)

- Mise en place d'un corpus de référence annoté en CR pour le français (où *RefGen* sera utilisé comme outil de pré-annotation de corpus)