PARSEME-FR: corpus d'évaluation annoté en EPs et ENs

Matthieu Constant, Marie Candito, Yannick Parmentier, Carlos Ramisch, Agata Savary

15 janvier 2018

Rappel PARSEME-FR WP1

EP: expression polylexicale
dont EN: entité nommée

Objectifs:

- Définir critères opérationnels pour définir et catégoriser les EPs
- Construire un corpus de référence annoté en EPs
  - de taille suffisante pour évaluation, pas forcément pour apprentissage
- Extraire et enrichir lexique d'après les données annotées

Corpus visé

- corpus Sequoia (Candito et Seddah 2012)
- 67000 tokens, 3099 phrases
- Textes de 4 origines: Europarl, Wikipedia, *L'Est Républicain*, rapports de l'Agence européenned du médicament (EMEA)
- Motivations:
  - licence libre
  - multiples couches d'annotations:
    - syntaxe, syntaxe profonde
    - (+mots composés grammaticaux uniquement)
- Prévu également: annotation sur une partie de French UD

En parallèle: shared task PARSEME

- Compétition PARSEME: identification des **E**Ps VERBALES (Savary et al., 17)
- guide d'annotation pour 18 langues
- pour le français: annotation 5000 VMWEs dans ce cadre
  - French UD corpus
  - corpus Sequoia

Annotation en EPs: Etat d'avancement

- double annotation terminée
  - sous Folia, cf. même outil que shared task PARSEME
  - pré-marquage automatique des EPs verbales annotées ds le cadre de la shared Task PARSEME
- adjudication: à moitié réalisée
  - script magique de Silvio
- Difficultés:
  - mise en place du guide: plus de temps que prévu
  - distinction/inclusion des EN dans les EP

Methodologie d'écriture du guide

Participants: Agata, Carlos, Marie, Mathieu, Yannick

- Pour chaque type d'EP, un responsable lit bibliographie (focus sur propriétés linguistiques)
- Première version de critères suffisants d'EP
- 4 Cycles, tant que l'accord n'est pas satisfaisant:
  - annotation pilote
  - adjudication, calcul d'accord
  - mise à jour des critères

Choix généraux: critères suffisants

- **Principe de base**: sauf exception, on définit des **critères suffisants**

→ Solution au problème bien connu de la **continuité** du statut d'EP

  - propriété non binaire

→ un critère d'EP suffit à l'annoter comme tel

→ enrichissement ultérieur du lexique extrait des annotations: ajout de tous les critères satisfaits

→ différentes vues sur les annotations pourront être obtenues par activation/inhibition de critères, selon les besoins

Choix généraux: ENs et EPs

Entités nommées également annotées

- Personnes, Lieux, Organisations, Produits, Evènements

Après moult débats internes: **ENs annotées séparément des EPs**

Frontière:

- dénomination conventionnelle d'une **entité spécifique**
- versus descripteur d'une **classe d'entités**, utilisé pour référence à entité spécifique ou pas, résolue en contexte
  - $\longrightarrow$ Frontière pas si claire
    - *le soleil*
    - *les Hectors ont plus confiance en eux que les Achilles*
    - *Association d'insertion des pays de la Saulx et du Perthois*

Choix généraux: ENs et EPs (suite)

- Cas facile: nom(s) propre(s)
  - en emploi standard, et pas pris comme une classe!
  - $\neq$ Un Jules ne peut pas avoir honte de son nom
- Cauchemar: ENs à base descriptive
  - *Association d'insertion des pays de la Saulx et du Perthois*
  - $\longrightarrow$ utilisation de ressources externes pour juger de la convention de nommage
- NB: une EN peut contenir une EP
  - *[[Chemin de fer]$_{EP}$ de l'Ouest]$_{ORG}$*

Critères d'identification des EPs: Objectif

Capture par critères $\approx$ formels de:

- la non compositionalité sémantique stricte
  - d'un point de vue applicatif: problématique pour l'analyse
  - *casser les pieds (à qqn)*
  - surplus de sens, en particulier termes: *juge d'instruction, traduction automatique*
- la variation réduite
  - d'un point de vue applicatif: problématique pour la génération
  - *cabine téléphonique* vs **cabine de téléphone*
  - *cabine téléphonique* vs **habitacle/pièce/chambre téléphonique*

Critères d'identification des EPs: Forme générale

Sauf exception, un critère de la forme:

- si l'application d'une modification "normalement licite ds ce contexte" crée
  - une inacceptabilité,
  - ou une modification de sens inattendue au regard de la modification apportée
- $\longrightarrow$ le critère est satisfait