Poster TALN-RECITAL

Détection automatique de chaînes de coréférence pour le français écrit :

règles et ressources adaptées au repérage de phénomènes linguistiques spécifiques

Bruno Oberle
LiLPa, Université de Strasbourg

Introduction et objectif

- Chaînes de coréférence : ensemble des expressions référentielles qui renvoient au même référent.
  Exemple: *Platon est un philosophe antique de la Grèce classique… Il reprit le travail philosophique de certains de ses prédécesseurs, notamment Socrate dont il fut l'élève.*
- Limites des systèmes existants: phénomènes spécifiques négligés:
  - anaphores infidèles *(mon chat… cet animal…)*
  - groupes *(Marie et Pierre)*
  - coréférence entre entités du discours indirect et pronoms de 1ère et 2e personnes dans le discours direct et les citations
  - anaphores zéro *(Pierre boit et ø fume)*
- **Objectif**: élaborer, pour le français écrit, un système end-to-end de détection automatique de la coréférence, qui s'attache à repérer des phénomènes coréférentiels fréquents dans les textes écrits mais qui sont souvent négligés par les autres systèmes.

Phénomènes spécifiques dans la résolution de la coréférence

ENTITÉS NOMMÉES

- Dictionnaire d'entités nommées contenant des informations nécessaires pour la coréférence:
  - genre (Marie Curie ⟶ elle (fém.), Ansel Adams ⟶ il (masc.))
  - reprises nominales possibles (Rhin ⟶ fleuve, cours d'eau…; Paris ⟶ ville, capitale…)
  - variation dans la désignation (Paris ⟶ Ville Lumière)
  - type de l'entité
- Méthode: Yago (Mahdisoltani *et al.*, 2014) (Wikipédia): pages de redirection, WordNet, balises

SYNONYMES ET HYPERONYMES

- Création d'un dictionnaire d'hyperonymes pour repérer les anaphores infidèles du type *mon chat… cet animal…*
- Méthode:
  - Glawi (Hathout et Sajous, 2016) (Wikitionnaire)
  - analyse des définitions dites par "genre prochain": un chat est *un mammifère carnivore*

PRONOMS DE 1ÈRE ET 2E PERSONNES

- Passage du discours indirect au discours direct (citations)
- Ex.: ***Paul** dit à Marie: "**Je** t'ai donné un cadeau."*
- Méthode: repérage *a priori* des citations (marques typographiques), des incises (… *dit-il*), détection de la coréférence par un ensemble de règles

Phénomènes spécifiques dans la détection des mentions

GROUPES

- Trois référents dans *Marie et Pierre*: Marie, Pierre, et le couple (groupe)
- Ex.: *Marie et Pierre sont heureux. Ils vont partir en vacances.*
- Méthode: Analyse syntaxique avec Talismane (Urieli, 2013), et recherche des groupes juxtaposés et/ou coordonnés avec des règles

ANAPHORES ZÉROS

- Sujets non exprimés, "en facteur commun": *Pierre boit et **ø** fume*
- Utile pour l'extraction d'informations
- Méthode: Analyse syntaxique avec Talismane (Urieli, 2013) et recherche des verbes juxtaposés ou coordonnés et "distribution" du sujet

Un système end-to-end

- **Repérage des mentions** en plusieurs passes:
  - repérage des entités nommées, du discours direct, extraction des incises dans les citations
  - tokenisation
  - analyse syntaxique avec Talismane (Urieli 2013)
  - correction de l'arbre résultant
  - ajout d'informations (groupes, anaphores zéro, etc.)
  - repérage des pronoms non référentiels
  - repérage des mentions

- **Détection de la coréférence**, plusieurs passes en fonction du type des relations et des informations nécessaires (syntaxiques ou sémantiques):
  - 1: anaphores liées (l'antécédent est repéré grâce à un calcul purement syntaxique)
  - 2: autres anaphores pronominales (pronoms et déterminants possessifs)
  - 3: entités nommées (fusion des chaînes partielles précédentes à l'aide des entités nommées)
  - 4: noms communs (fusion des chaînes partielles sur la base du dictionnaire d'hyperonymes)

Évaluation

- Nous proposons une évaluation provisoire sur un petit corpus annoté spécifiquement (littérature, faits divers, textes de FLE):
- Comparaison avec RefGen (Longo 2013), le système à base de règles qui se rapproche le plus de nôtre.

	identification	MUC	B3	CEAF	BLANC
notre système	83.62	58.1	64.53	69.28	53.08
RefGen	50.2	36	31.9	29.9	24.9

Conclusion

- Apport:
  - développement de ressources linguistiques pour la résolution de la coréférence (anaphores infidèles)
  - règles permettant de repérer des phénomènes linguistiques spécifiques (groupes, anaphores zéro, pronoms de 1ère et 2e personnes dans les citations)
  - système end-to-end à base de règles
- Perspectives:
  - intégration des règles, en seconde passe, dans un système à base d'apprentissage automatique (hybridation)

Références

- Hathout N. & Sajous F. (2016). *Wiktionnaire's wikicode glawified : a workable French machine-readable dictionary.* LREC'16.
- Longo L. (2013). *Vers des moteurs de recherche "intelligents": un outil de détection automatique de thèmes.* Thèse.
- Mahdisoltani F., Biega J. & Suchanek F. (2014). *Yago3 : A knowledge base from multilingual wikipedias.* 7th Biennial Conference on Innovative Data Systems Research.
- Urieli A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit.* Thèse.

Remerciement

Ce travail a été réalisé avec le soutien du projet ANR Democrat ("Description et modélisation des chaînes de référence : outils pour l'annotation et le traitement automatique", ANR-15-CE38-0008).