Détection automatique de mentions dans un corpus de français oral

Loïc Grobol[1], Isabelle Tellier[1], Éric de la Clergerie[2], Marco Dinarelli[1] & Frédéric Landragin[1]

(1) Lattice (CNRS, ENS Paris, Université Sorbonne Nouvelle, PSL Research University, USPC) (2) ALMAnaCH (Inria)

Mentions et entités pour la coréférence

On appelle MENTION toute expression faisant référence à une ENTITÉ du discours.

Mentions

[je] viens de voir [quelqu'un de [la mairie d' [Orléans]]] euh comme [je] connais [quelqu'un] [je] vais réussir à [me] faire réparer [deux ou trois trous]

Entités

«je» «je» «je» «me» | «quelqu'un de la mairie d'Orléans» «quelqu'un» | «la mairie d'Orléans» | «Orléans» | «deux ou trois trous»

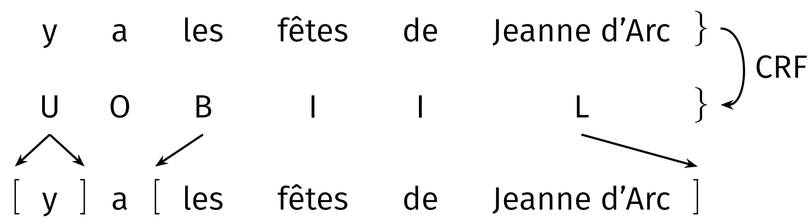
Tâche Détecter les **mentions** dans du texte brut (étape 1.)

- Pour l'étape 2: CROC (Désoyer et al. 2015) détecte les chaînes de coréférence mais ne détecte pas les mentions.

Corpus ANCOR (Muzerelle et al. 2013).

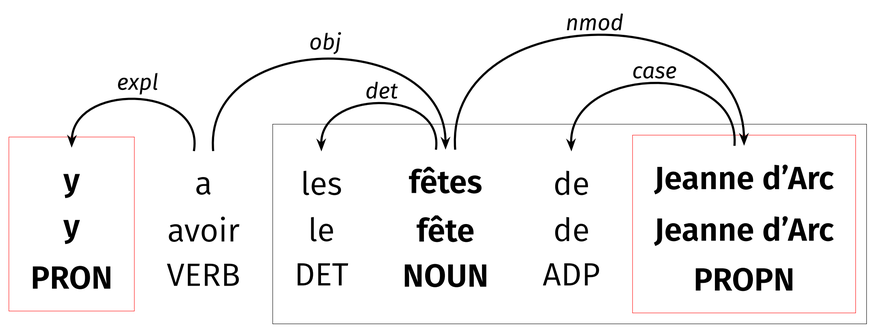
- Français oral transcrit
- Seul corpus d'**oral** annoté en chaînes de coréférences toutes langues confondues

Apprentissage automatique



- Étiquetage de séquences par CRF linéaire (Wapiti, Lavergne et al. 2010)
  - Schéma BILOU (*Begin*, *Inside*, *Last*, *Outside*, *Unique*)
  - Pas d'enchâssement: uniquement les mentions les plus larges
  - Traits utilisés: forme, lemme, POS, chunks, EN (SEM, Tellier et al. 2012)
- Apprentissage sur ANCOR train/dev/test arbitraire: 80 %/10 %/10 %
- Raisonnablement satisfaisant par rapport à l'état de l'art, mais comparaison délicate
  - Les systèmes de NER tiennent en général compte d'un typage
  - Les systèmes de détection de coréférence tiennent souvent compte des enchâssements, ou ne tiennent pas compte des singletons

Analyse syntaxique



- Analyses syntaxiques en dépendances: Talismane (Urieli 2013)
- Candidats-mentions comme pseudo-constituants reconstruits
  - Repérer les **têtes** à partir de leurs *part-of-speech*
  - Suivre les dépendances (pour les noms)
- Pour l'instant, pas de filtrage des non-référentiels
- Rappel comparable à l'état de l'art, précision largement améliorable (mais une large partie de l'état de l'art utilise des analyses syntaxiques *gold*).

Résultats

Comparaison avec l'état de l'art pour la détection de mentions. Résultats à considérer avec prudence compte-tenu de l'hétérogénéité des conventions d'annotation et d'évaluation.

Système	Corpus	$P(\%)$	$R(\%)$	$F_1(\%)$	Méthode
Lassalle 2015	CoNLL (eng)[1]	43,77	97,97	60,05	Parse[3, 2]
Kummerfeld et al. 2011	OntoNotes (eng)[1]	56,97	69,77	62,72	Parse[2]
Ogrodniczuk et al. 2014	PCC (pol)	66,04	63,99	65,00	Chunk
Soraluze et al. 2012	EPEC (eus)	76,85	78,59	77,58	Chunk
Uryupina et al. 2013	CoNLL (ara)[1, 5]	66,0	66,1	66,1	Parse[4, 2]
Uryupina et al. 2013	CoNLL (zho)[1, 5]	68,9	71,3	70,1	Parse[4, 2]
Uryupina et al. 2013	CoNLL (ara)[1, 5]	31,07	90,67	46,28	Parse[3, 2]
Nguyen et al. 2016	ACE 2005 (eng)[5]	83,7[6]	81,8[6]	82,7[6]	Direct
Nos expériences	**ANCOR**	**57,28**	**77,07**	**65,72**	**Parse**
Nos expériences	**ANCOR[5]**	**89,09**	**88,61**	**88,85**	**Direct**

1. Ne tient pas compte des singletons
2. Utilise une analyse syntaxique de référence
3. Optimisé pour le rappel
4. Optimisé pour la précision
5. Sans mentions enchâssées
6. Tient compte de la catégorisation des mentions

Perspectives

- Utiliser des ressources adaptées pour l'oral
- Combiner plusieurs méthodes dans un système hybride
  - Détecter les mentions maximales par apprentissage direct
  - Détecter les sous-mentions avec une analyse syntaxique
- Filter les expressions non-référentielles
- Développer un système *end-to-end*
  - S'interfacer avec CROC (Désoyer et al. 2015)
  - Traiter détection des mentions et constructions des chaînes comme une tâche jointe

Références

Nguyen, Thien Huu et al. (2016). "Toward Mention Detection Robustness with Recurrent Neural Networks". In: *CoRR* abs/1602.07749.

Désoyer, Adèle et al. (2015). "Coreference Resolution for Oral Corpus: a machine learning experiment with ANCOR corpus". In: *Traitement Automatique des Langues* 55.2.

Lassalle, Emmanuel (2015). "Structured learning with latent trees: a joint approach to coreference resolution". PhD thesis. Université Paris Diderot Paris 7.

Ogrodniczuk, Maciej et al. (2014). "Detection of Nested Mentions for Coreference Resolution in Polish". In: *Advances in Natural Language Processing: 9th International Conference on NLP*.

Muzerelle, Judith et al. (2013). "ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement". In: *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles*.

Urieli, Assaf (2013). "Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit". PhD thesis. Université de Toulouse II le Mirail.

Uryupina, Olga et al. (2013). "Multilingual Mention Detection for Coreference Resolution." In: *IJCNLP*.

Soraluze, Ander et al. (2012). "Mention detection: First steps in the development of a Basque coreference resolution system". In: *Proceedings of KONVENS 2012*.

Tellier, Isabelle et al. (2012). "Apprentissage automatique d'un chunker pour le français". In: *Actes de la 19e conférence sur le Traitement Automatique des Langues Naturelles*.

Kummerfeld, Jonathan K. et al. (2011). "Mention Detection: Heuristics for the OntoNotes Annotations". In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*.

Lavergne, Thomas et al. (2010). "Practical Very Large Scale CRFs". In: *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*.