DEMOCRAT : description et modélisation des chaînes de référence

Frédéric Landragin, Isabelle Tellier, Yoann Dupont

Langues, Textes, Traitements Informatiques, Cognition

UMR 8094

Comme tout avait brûlé – la mère, les meubles et les photographies de la mère -, pour Fabre et le fils Paul c'était tout de suite beaucoup d'ouvrage : toute cette cendre et ce deuil, déménager, courir se refaire dans les grandes surfaces. Fabre trouva trop vite quelque chose de moins vaste, deux pièces aux fonctions permutables sous une cheminée de brique dont l'ombre donnait l'heure, et qui avaient ceci de bien d'être assez proches du quai de Valmy.

Le soir après le dîner, Fabre parlait à Paul de sa mère, sa mère à lui Paul, parfois dès le dîner. Comme on ne possédait plus de représentation de Sylvie Fabre, il s'épuisait à vouloir la décrire toujours plus exactement : au milieu de la cuisine naquirent des hologrammes que dégonflait la moindre imprécision. Ça ne se rend pas, soupirait Fabre en posant une main sur sa tête, sur ses yeux, et le découragement l'endormait. Souvent ce fut à Paul de déplier le canapé convertible, transformant les choses en chambre à coucher.

Description et modélisation des chaînes de références
Outils pour l'annotation de corpus et le traitement automatique

Frédéric Landragin, Isabelle Tellier, Yoann Dupont et le consortium *Democrat*

Le projet ANR Democrat vise à développer les recherches sur la langue et la structuration textuelle du français via l'analyse détaillée et contrastive des chaînes de références (instanciations successives d'une même entité) dans un corpus diachronique de textes écrits entre le 9ème et le 21ème siècle, avec des genres textuels variés. Il réunit des chercheurs issus des laboratoires Lattice, LiLPa et ICAR. Il a été lancé en mars 2016 et l'essentiel des efforts porte actuellement sur la constitution et l'annotation (manuelle) d'un corpus. Plusieurs expérimentations d'annotation ont eu lieu, de manière à tester différentes procédures, et notamment avec ou sans pré-annotation (automatique) des expressions référentielles. Ce poster décrit cette première phase de travaux.

Objectifs et livrables du projet

Proposer un modèle de la référence et de la composition des chaînes de références
- modèle orienté sur le discours et pas seulement la phrase
- modèle qui s'enrichit de comparaisons inter-langues et d'études diachroniques
- perspectives : étude des transitions référentielles et de la saillance référentielle

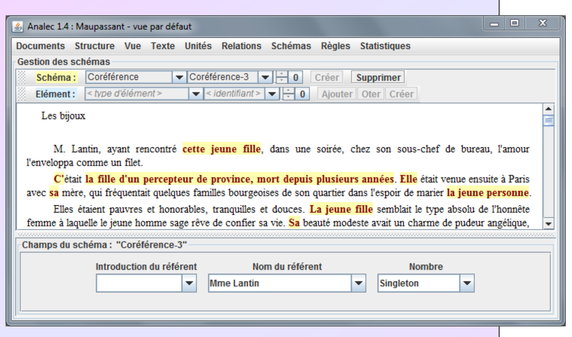
Fournir un corpus annoté qui serve de corpus de référence et d'apprentissage
- taille visée : 1 million de mots, 100.000 maillons de chaîne annotés
- proposer un pendant au seul corpus similaire existant pour le français : ANCOR

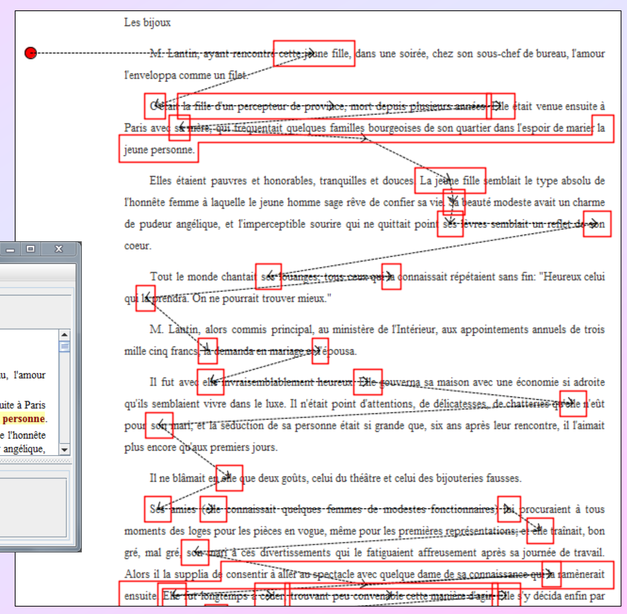
Développer un outil d'annotation adapté aux chaînes de références
- prototype de départ : ANALEC
- intégration des fonctionnalités d'annotation et de gestion de schémas d'annotation dans TXM

Développer un système de résolution automatique de la coréférence
- techniques d'apprentissage artificiel (SVM, CRF) appliquées sur le corpus annoté manuellement
- participation envisagée à une campagne d'évaluation internationale

Annotation manuelle du corpus *Democrat*

Tests avec GLOZZ comme avec ANALEC :
- unités de type « *mention* »
- schémas « *chaîne de coréférences* »

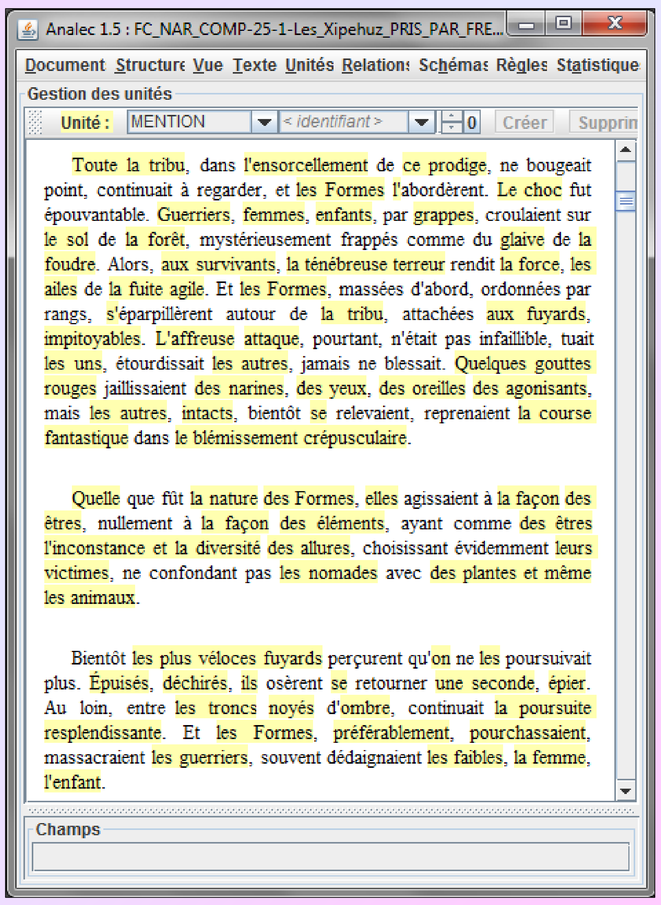




Pré-annotation automatique : exploitation de SEM

Utilisation de SEM, en mode « *chunker nominal* », avec utilisation de LEFFF, génération d'étiquettes POS et de catégories d'entités nommées

17	Comme	CS	O	O
18	il	CLS	B	O
19	faisait	V	O	O
20	une	DET	B	O
21	chaleur	NC	I	O
22	de	P	O	O
23	33	DET	B	O
24	degrÃ©s	NC	I	O
25	,	PONCT	O	O
26	le	DET	B	O
27	boulevard	NC	I	O
28	Bourdon	NPP	B	O
29	se	CLR	B	O
30	trouvait	V	O	O
31	absolument	ADV	O	O
32	dÃ©sert	ADJ	O	O
33	.	PONCT	O	O
34				
35	Plus	ADV	O	O
36	bas	ADJ	O	O
37	,	PONCT	O	O
38	le	DET	B	O
39	canal	NPP	I	O
40	Saint-Martin	NPP	I	B-Location
41	,	PONCT	O	O
42	fermÃ©	VPP	O	O
43	par	P	O	O
44	les	DET	B	O
45	deux	ADJ	I	O
46	Ã©cluses	NC	I	O
47	,	PONCT	O	O
48	Ã©talait	V	O	O



Chunks nominaux, entités nommées et expressions référentielles

Chunks nominaux
- tous les chunks nominaux d'un texte ne réfèrent pas et ne participent donc à aucune chaîne
- un pré-repérage automatique des chunks nominaux permet à l'annotateur de ne rien oublier
- un tel pré-repérage nécessite d'enlever le « bruit », par exemple les « il » impersonnels (qui sont repérés par SEM) et les mentions non référentielles de parties du corps (« prendre la tête », etc.)
- un chunker, par définition, repère des portions de texte non enchâssées ; or des compléments du nom tels que « l'introduction du rapport du comité d'évaluation » enchâsse plusieurs références, qu'il s'agit donc de reconstruire à partir des chunks

Entités nommées
- exploiter un reconnaisseur d'entités nommées permet de catégoriser quelques chunks : références à des institutions, des lieux, etc.
- c'est une propriété qui s'ajoute aux chaînes de références et non aux expressions référentielles

Expressions référentielles et choix d'annotation pour le corpus du projet Democrat
- phase 1 : depuis le texte brut, on exploite le chunker nominal SEM en tant que pré-annotation
- phase 2 : à la main, dans ANALEC, on corrige les chunks pour délimiter les expressions référentielles, c'est-à-dire qu'on supprime le bruit, qu'on crée des expressions nouvelles qui ne correspondent à aucun chunk, qu'on corrige les bornes de certains chunks
- phase 3 : après cette délimitation, on remplit une propriété (que SEM n'a pas pu remplir), celle comportant un identifiant du référent
- phase 4 : une fois tout le texte annoté, ANALEC génère automatiquement des schémas, un pour chaque chaîne de coréférences

Références

Désoyer, A., Landragin, F., Tellier, I., Lefeuvre, A., Antoine, J.-Y. (2014) « Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR », *TAL*, 55(2), pp. 97-121.
Dupont Y., Tellier I. (2014) « Un reconnaisseur d'entités nommées du Français », *Traitement Automatique des Langues Naturelles (TALN 2014)*, session démo, Marseille.
Landragin F. (2016) « Conception d'un outil de visualisation et d'exploration de chaînes de coréférences », *Thirteen International Conference on Statistical Analysis of Textual Data (JADT 2016)*, Nice.
Landragin F., Poibeau T., Victorri B. (2012) "ANALEC: a New Tool for the Dynamic Annotation of Textual Data", *LREC 2012*, Istanbul, Turquey.
Tellier I., Eshkol I., Dupont Y., Wang I. (2014) « Peut-on bien chunker avec de mauvaises étiquettes POS ? », *Traitement Automatique des Langues Naturelles (TALN 2014)*, Marseille.
http://www.agence-nationale-recherche.fr/?Projet=ANR-15-CE38-0008

Travail réalisé avec le soutien de l'ANR dans le cadre du projet *Democrat* ANR-15-CE38-0008