Reconnaissance d'Entités Nommées par Extraction Automatique de Transducteurs

Damien Nouvel, Jean-Yves Antoine, Nathalie Friburger, Arnaud Soulet

Université François-Rabelais Tours — Laboratoire d'Informatique EA 2101

Extraction de motifs pour la REN

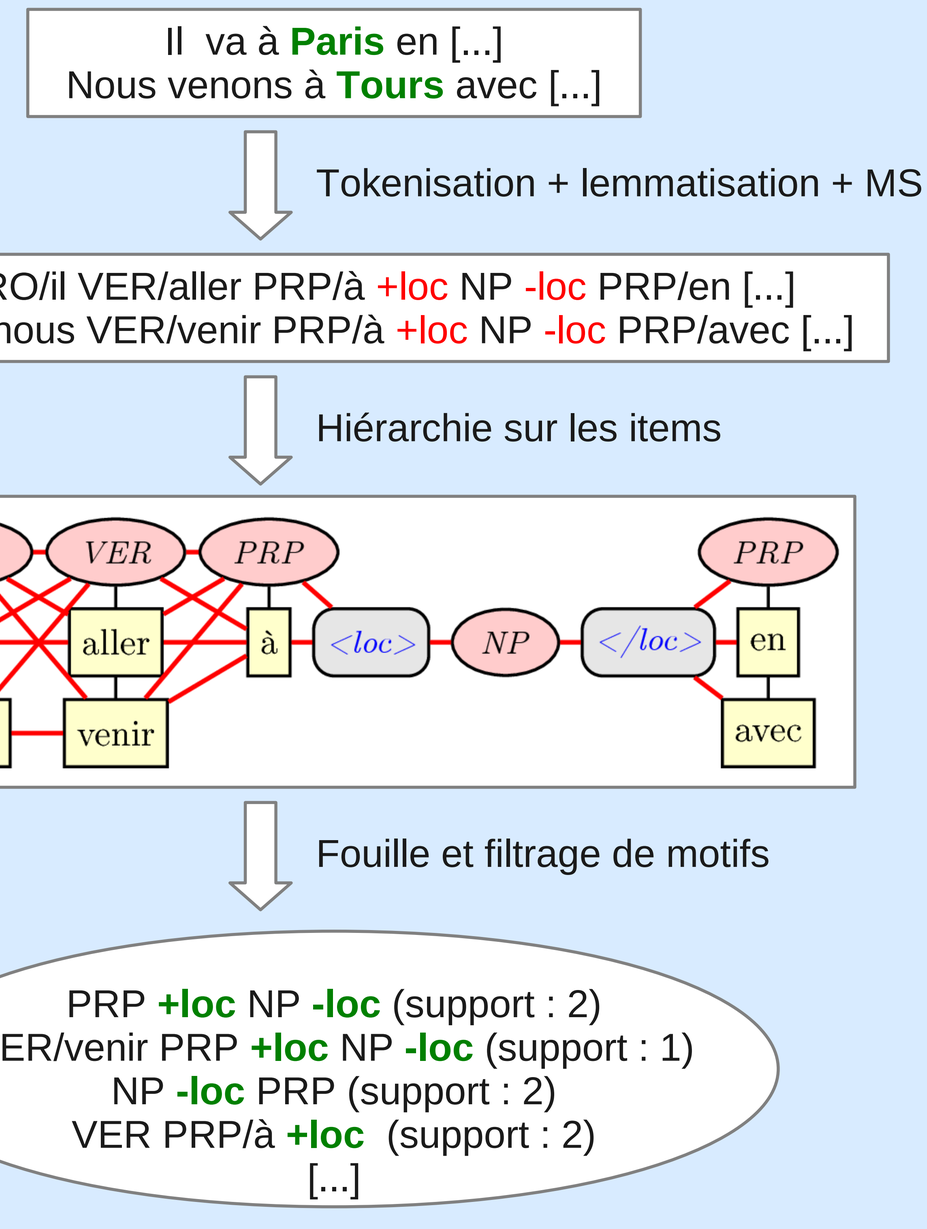
Extraction d'information par combinaison d'approches :
- **symboliques** (transducteurs sous Unitex : CasEN),
- **apprentissage** automatique (MaxEnt, HMM, CRF).

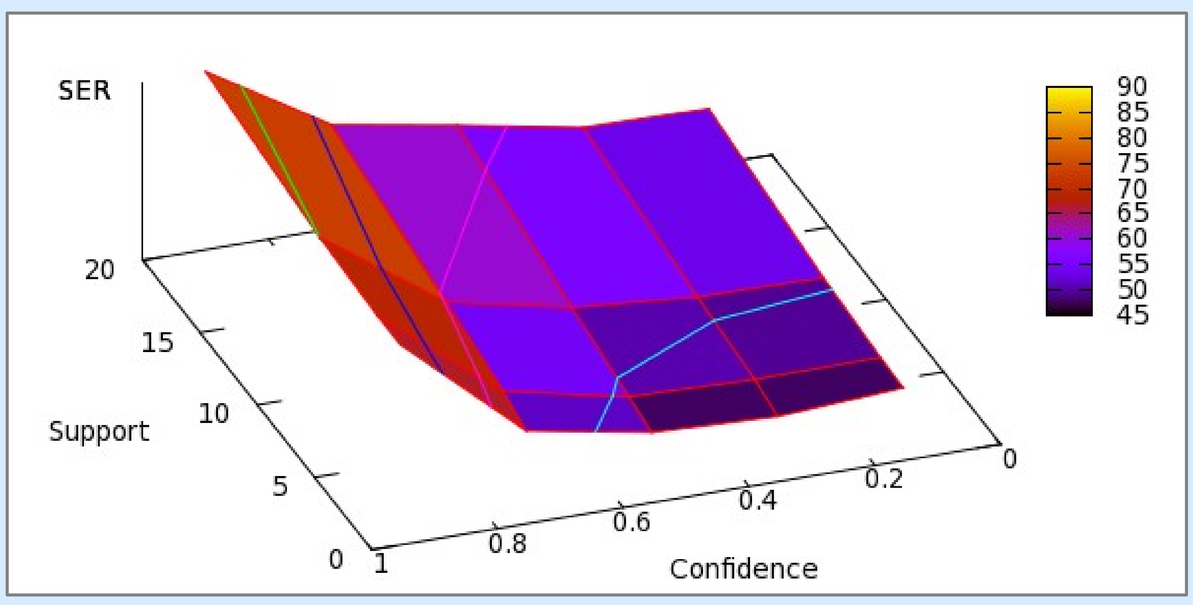
Prétraitements linguistiques par **étiquetage morpho-syntaxique** :
- **tokenisation** et **lemmatisation** des données,
- **segmentation** en phrases,
- construction d'une **hiérarchie** (taxonomie) sur les items,
- suppression des items lexicaux pour les noms propres (NP).

Fouille de données :
- **séquentielle** (phrases = séquences, tokens = items),
- supervisée : les frontières d'EN sont des items, les **marques**,
- recherche exhaustive (seuils de support et confiance) des **motifs contigus** qui comportent au moins une marque (complets ou « partiels »).

Filtrage des motifs redondants, pour une même fréquence / confiance :
- **maximaux** (plus longs / spécifiques pour la hiérarchie),
- **informatifs** (qui insèrent le plus de marques possibles).

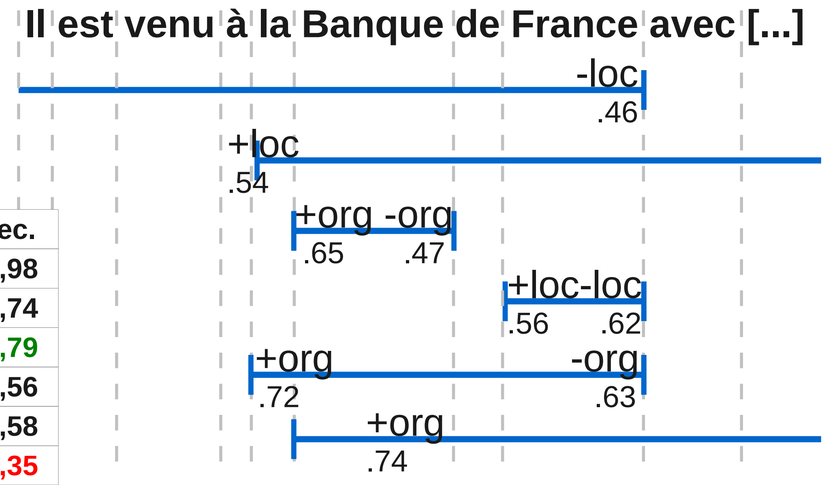
Cadre d'application, **Ester2** :
- transcriptions manuelles (disfluences, bruit limité),
- corpus fouille / test : 45Kmots pour validation croisée à 12 plis,
- proportion moyenne d'EN : 7 / 100 tokens,
- taxonomie Ester2 (pers, loc, org, amount, time, fonc, prod).





Un modèle discriminant par « marques »

Modèle à **maximum d'entropie** :
- les motifs sont les « features » du modèle, pour l'apprentissage (corpus Ester2, 50Kmots) des **probabilité des marques**,
- l'insertion d'une marque est interprétée comme une probabilité de **transition** pour réaliser un **étiquetage** (plat) en **entités nommées**,
- annotation par un algorithme de **programmation dynamique**.



R \ P	=	+pers	-pers	+loc	-loc	+org	-org	+fonc	-fonc	+time	-time	+amo	-amo	total	rec.
=	**27168**	46	5	114	68	91	75	28	28	77	76	14	13	27803	**0,98**
+pers	86	**430**		20	1	26	1		18	1				583	**0,74**
-pers	48		**470**		45		27			1	1			592	**0,79**
+loc	162	20	2	**394**		114	1		2	3	2			700	**0,56**
-loc	137	2	16	2	**407**		127			4	3			698	**0,58**
+org	203	30		45		**157**		2	6	3	2			448	**0,35**
-org	176		59		69		**122**		2	5	8	2		443	**0,28**
+fonc	84	1	2	3		2		**129**		4				225	**0,57**
-fonc	112	27	6		10		14		**48**	1	1			219	**0,22**
+time	249	2	4	1	12		4			**223**	4	8	1	508	**0,44**
-time	200	1			6		2			2	**293**	1	12	517	**0,57**
+amo	98			1	1					6	2	**21**	1	130	**0,16**
-amo	79		1				1				17		**35**	133	**0,26**
total	28802	559	565	580	6.19	390	374	159	104	330	409	46	62		
prec.	**0,94**	**0,77**	**0,83**	**0,68**	**0,66**	**0,4**	**0,33**	**0,81**	**0,46**	**0,68**	**0,72**	**0,46**	**0,56**		

R \ P	=	+	-	total	rec.
=	**27168**	370	265	27803	**0,98**
+	882	**1646**	66	2594	**0,63**
-	752	48	**1802**	2602	**0,69**
total	28802	2064	2133		
prec.	**0,94**	**0,8**	**0,84**		

Discussions et perspectives

→ Performances pas (encore ?) au rendez-vous : SER ~46%, précision ~72%, rappel ~53%
→ Dépendance des modèles aux données d'entrainement fouille / MaxEnt (Ester2) et importance de leur adéquation vis-à-vis des données de test
→ Hybridation / intégration des motifs fouillés avec un système symbolique (par ex. transducteurs / CasEN)
→ Enrichissement des motifs (dictionnaires, exp. composées, cat. verbale, chunking) pour mieux déterminer les marques (org / pers / loc, time / amount),
→ Amélioration de l'annotation en entités nommées à partir des probabilités de marques (HMM, CRF, récursivité, etc.),

Ce travail a été réalisé dans le cadre des projets *Variling* (ANR-06-CORP-023), *Epac* (ANR-00-MDCA-006-03), créés par l'*Agence Nationale de la Recherche* (ANR) et FEDER *Région Centre*.