Annotation d'Entités Nommées par Extraction de Règles de Transduction

Damien Nouvel et Arnaud Soulet
prenom.nom@univ-tours.fr

Université François Rabelais Tours
Laboratoire d'Informatique
Equipe BDTLN

Plan de la présentation

1. La Reconnaissance des Entité Nommées (REN)
2. Extraction de règles de transduction
3. Application des règles pour une solution d'annotation
4. Conclusion

Plan de la présentation

Les Entités Nommées (EN)

- **Objets linguistiques** pour la **recherche d'information**...
- **Noms propres** :
  - Personnes, lieux
  - Organisations (sociétés, administrations...)
  - Productions humaines (voitures, livres, portables, films...)
- **Descriptions définies** :
  - Expressions de temps
  - Montants / quantités (physiques, monétaires...)
  - Fonctions (emploi, occupation, position)

Exemple

L'*iPhone 4* a été annoncé à la conférence *du 7 juin 2010* par *Steve Jobs*, le *dirigeant* américain de la compagnie *Apple*. Il pèse *140g*.

Approches

La tâche de Reconnaissance des Entités Nommées

- **Détecter** / délimiter les EN (bornes, extension, frontières)
- **Catégoriser** / classifier les EN (types, classes, traits)

Systèmes de REN

- **Symboliques** : règles de reconnaissance, souvent implémentés manuellement (CasEN...)

$\Rightarrow$ précis, mais faire développer un système qui soit suffisamment couvrant est assez **coûteux**

- **Apprentissage automatique** : CRF, Bayes, Entropie, HMM...

$\Rightarrow$ assez bon rapport performance / coût de développement, mais difficiles à exploiter et à faire évoluer en tant que **ressource**

Notre approche : motivation

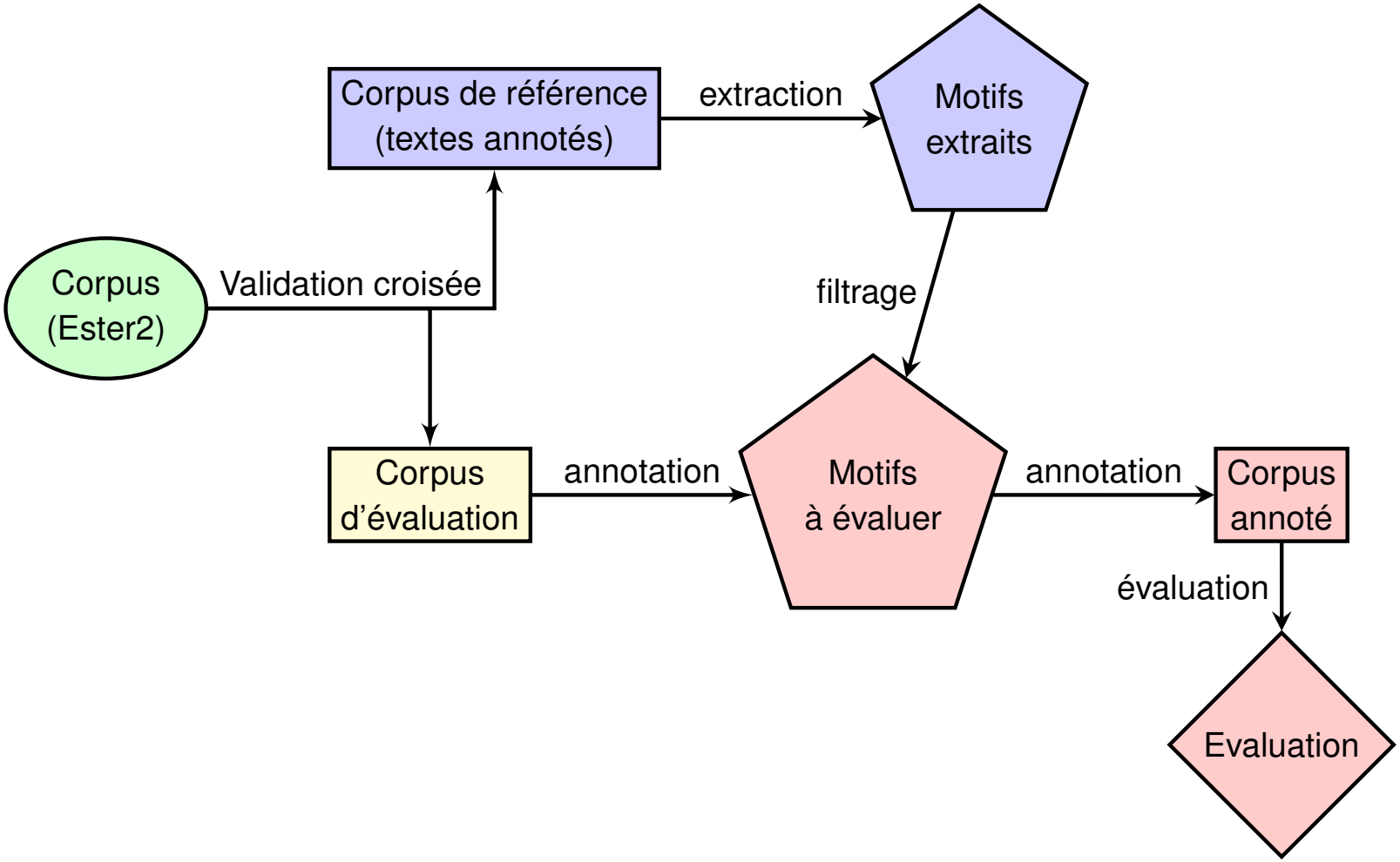
Exemples (lieux)

(. . .) il a prévu d'*aller à Brest* demain pour une conférence (. . .)
(. . .) nous aimerions nous *rendre à Barcelone* prochainement (. . .)
(. . .) son supérieur l'a *envoyé à Londres* pour qu'il (. . .)
$\Rightarrow$ **verbe + préposition *à* + X** $\rightarrow$ *<loc>* **X** *</loc>*

Découverte de règles pour la REN

- Quels types de **motifs** pour constituer ou améliorer une **base de connaissances** REN ?
- A partir de textes annotés (de référence), comment **extraire** des motifs comme **règles symboliques** ?
- Comment **évaluer** la qualité des règles et **annoter** un texte à l'aide de ces règles ?

Notre approche : méthodologie



Plan de la présentation

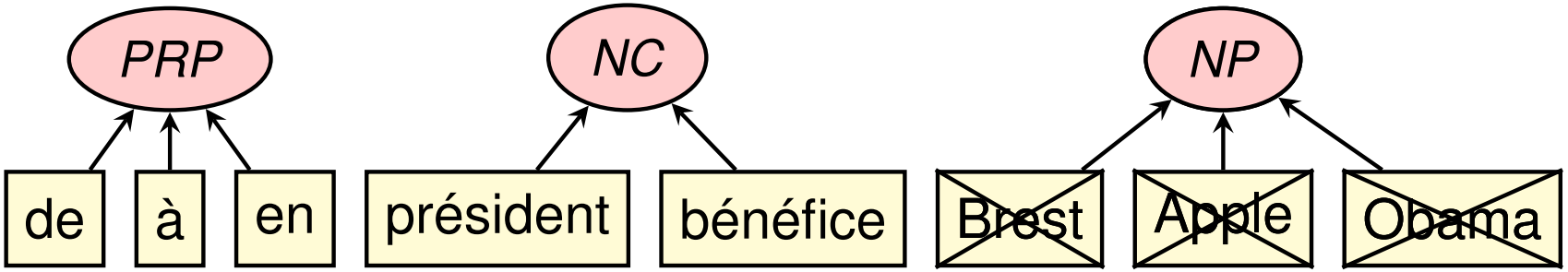
Règles de transduction : des motifs pour la REN

- Objectif : trouver des **règles** qui permettent d'**insérer** des **marques de début** / **fin** d'un **type** d'EN
  - il *va à* Brest $\rightarrow$ il va à *<loc>* Brest *</loc>*
  - le *président* Obama $\rightarrow$ le président *<pers>* Obama *</pers>*
  - le *bénéfice* d'Apple a été $\rightarrow$ le bénéfice d' *<org>* Apple *</org>* a été
- Utilisation de techniques de **fouille de données**
  - **Recherche** de motifs sur des textes de référence (corpus annotés)
  - **Extraction** des motifs **corrélés aux marques d'EN**
  - Pouvoir **généraliser** les motifs extraits
- Les noms propres, une catégorie **"ouverte"**
  - Nombreux et évolutifs (apparition quotidienne de nouvelles formes)

  $\Rightarrow$ Ne pas s'appuyer sur les noms propres (Brest, Obama, ...)

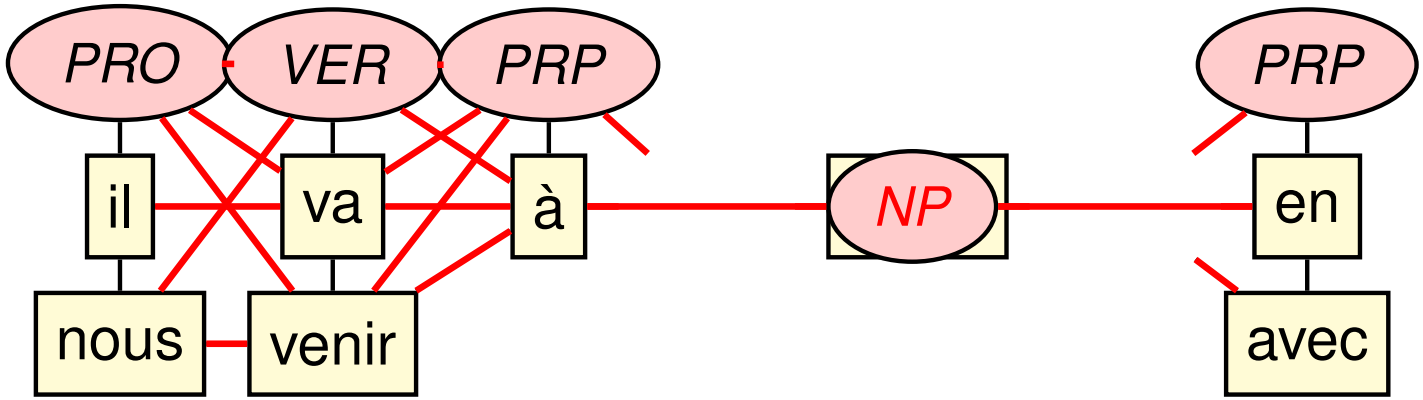
Prétraitements linguistiques du corpus

- Un outil (TreeTagger) nous permet de réaliser simultanément :
  - **Tokenization** : segmentation en “mots” (items)
  - **Lemmatisation** : trouver la forme “normale” du mot
  - **Catégorisation morpho-syntaxique** : “rôle grammatical” (Nom Commun, Nom Propre, VERbe, PRePosition, ponctuation etc.)
- Catégories morpho-syntaxiques :
  - Segmentation en **“phrases”** ($\simeq$ **séquences**)
  - **Ordre partiel** (hiérarchie, forêt) sur les lemmes (items)
  - Noms Propres : on ne conserve **pas** le mot



$\Rightarrow$ Extraction des motifs contigüs (aux marques EN près)

Du corpus aux motifs



- (. . .) Et comme il *va à Brest* en train, il se dit (. . .)
- (. . .) , pour une fois nous *viendrons à Barcelone* avec (. . .)

Motifs hiérarchiques extraits

-
-
-
-
- *il aller à <loc> NP </loc> en*

Exemple : règles de transduction

Texte fouillé

Il est allé à *<loc>* Paris *</loc>* hier. Etant venu de *<loc>* Brest *</loc>* dans l'après-midi, il arriva en soirée. Il alla chez *<pers>* Marie *</pers>*, puis il prit le métro en direction de *<loc>* Clichy *</loc>*, et s'arrêta à *<loc>* Saint Lazare *</loc>* pour aller à pied récupérer un colis, qui venait d'*<loc>* Ecosse *</loc>*. (. . .)

Règles de transduction extraites

Règle de transduction	Fréquence	Confiance
venir PRP <loc> NP </loc>	2	2/2
aller PRP <loc>	3	1/3
PRP <loc>	10	5/10
de <loc> NP </loc>	2	2/2

Règles de transduction

Règles de transduction

Une **règle de transduction** est un **motif morpho-syntaxique** (reposant sur la hiérarchie des types morpho-syntaxiques) auquel est associée **la confiance** avec laquelle nous pouvons l'appliquer comme **transducteur** afin d'insérer des **marques** dans un texte.

Règles (de transduction) informatives

- Fouille exhaustive : nombre de règles extraites **très important**, beaucoup de redondance

$\Rightarrow$ Nécessité de **filtrer** les règles : pour deux règles, dont l'une est spécialisation de l'autre, de même fréquence / confiance

  - Plus **générale** (en utilisant la hiérarchie morpho-syntaxique)
  - Plus **informative** (nombre de marques insérées)

Le corpus Ester2

Caractéristiques du corpus

- Ensemble d'émissions **radiodiffusées** (2007-2008)
- 12 fichiers
  - **1 300** séquences
  - **40 000** mots
  - **2 798** entités nommées
- Transcriptions **manuelles**, mais avec les difficultés de l'oral :
  - Phrases longues, répétitions, hésitations, tours de parole, etc.

  $\Rightarrow$ Lemmatisation et catégorisation **moins performantes**

  $\Rightarrow$ Plus de **bruit** pour extraire des règles

Extraction de règles sur Ester2

Implémentation de l'extraction

- Algorithme d'extraction de motifs séquentiels **par niveaux** avec **élagage** par seuil de fréquence
- Mise à profit de la hiérarchie pour accélérer l'extraction

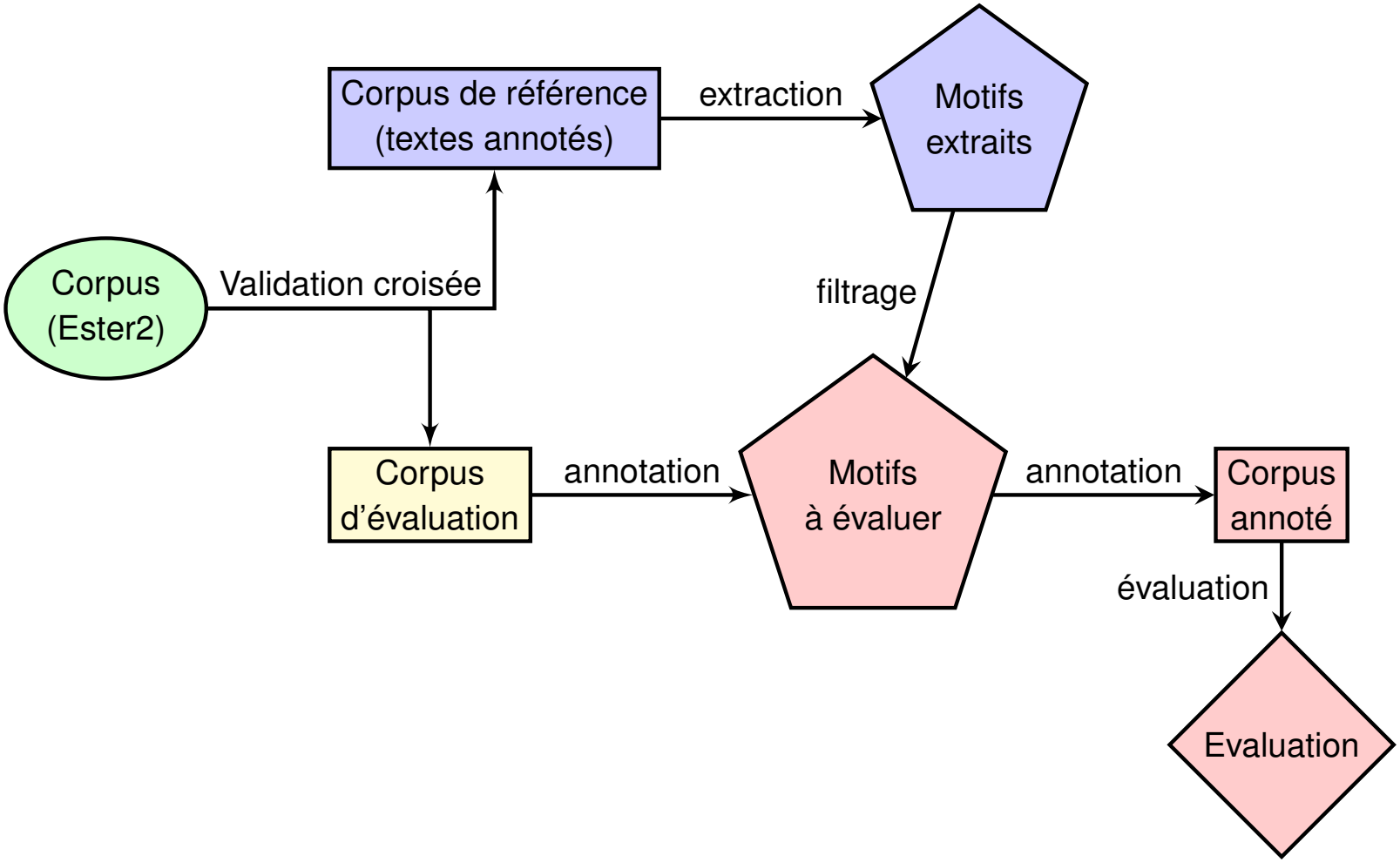
Quantité de règles extraites

F.	C.	Règles	Règles informatives	Temps
3	.45	66 961	11 922	0'19"
5	.60	13 676	3 679	0'7"
9	.45	3 380	1 742	0'3"

TABLE: Extraction de règles sur Ester2 à seuils de fréquence (F.) et de confiance (C.) (CPU 2.4GHz, RAM 4Go)

Plan de la présentation

Notre approche : méthodologie



Annotation par confiance

Règles complètes

- Comportent une marque de début et une marque de fin d'EN
- Appliquer les règles tant que la portion qu'elles reconnaissent n'a pas déjà été annotée

$\Rightarrow$ Par ex. :

- *le président <pers> NP NP </pers>*
- *aller à <loc> NP </loc>*

Avantages

- Utilisation de la **confiance**
- Algorithme facile à implémenter et rapide

Inconvénients

- **Sous-ensemble** des règles extraites
- Ordre par **confiance** des règles ?

Résultats expérimentaux : règles complètes

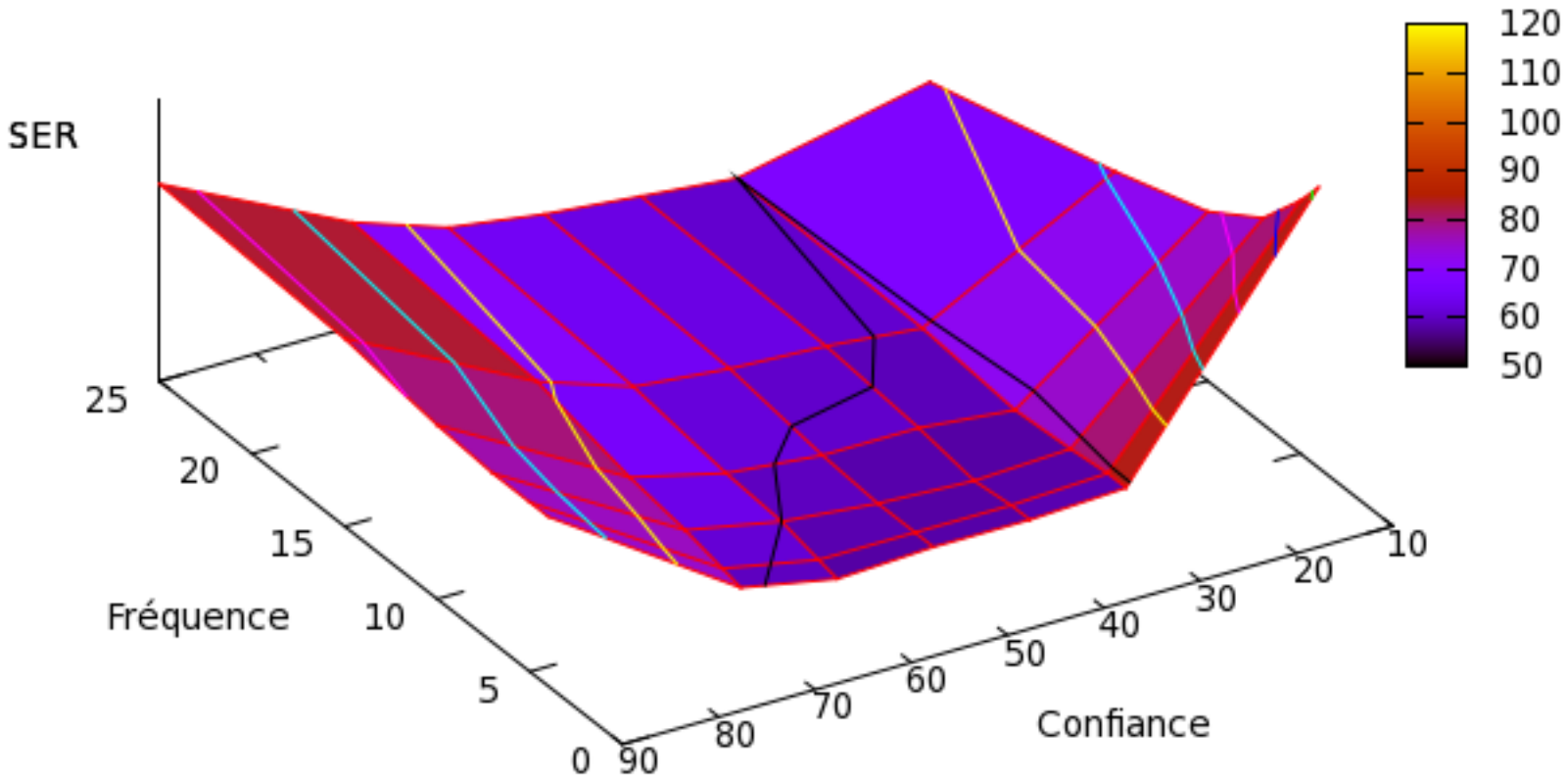


FIGURE: Evaluation (SER, à minimiser) des annotations produites sur Ester2 à seuils de fréquence et confiance

Stratégie d'annotation par combinaison de règles

Règles partielles

- N'importe quelles marques, dans n'importe quel ordre
- **Exploiter la confiance des règles** / **marques** pour une annotation probable et valide (programmation **dynamique**)

$\Rightarrow$ Par ex. :

- *DET* président *<pers> NP*
- la Banque de *NP </org> VER*

Avantages

- Exploitation de **toutes** les règles
- Possibilité de **plusieurs règles** par marque

Inconvénients

- Grand espace des **solutions**
- Calcul des **probabilités** par marques ?

Résultats expérimentaux : règles partielles

Paramètres		Règles complètes				Règles partielles			
F.	**C.**	**P..**	**R.**	**F-m**	**SER**	**P..**	**R.**	**F-m.**	**SER**
3	.40	67,66	50,19	0,58	55,14	62,91	57,29	0,6	52,96
3	.45	70,66	48,37	0,57	**54,82**	65,91	54,76	0,6	**51,99**
3	.60	76,61	45,52	0,57	55,66	72,53	50,62	0,6	53,66
5	.40	69,63	45,87	0,55	56,75	65,41	52,08	0,58	54,09
5	.45	72,01	44,39	0,55	56,91	67,52	49,91	0,57	54,45
5	.60	78,9	41,49	0,54	57,72	76,73	45,05	0,57	56,42
9	.40	73,9	40,89	0,53	57,78	69,78	46,41	0,56	54,9
9	.45	76,14	40,4	0,53	57,58	73,62	44,92	0,56	55,42
9	.60	81,49	37,41	0,51	60,92	80,04	40,72	0,54	58,62

TABLE: Comparaison règles complètes / partielles (fréquence F., confiance C., précision P., rappel R., f-mesure F-m., SER)

Plan de la présentation

Conclusion

Contributions

- **Extraction** de règles sur une hiérarchie (**morpho-syntaxique**)
- Filtrage des motifs **généralisés** et **informatifs** comme **règles**
- Stratégie pour **annoter** à partir de **règles partielles**

Difficultés

- Mieux **filtrer** les motifs pour **une base de connaissances** ?
- **Exploiter toutes les règles** pour réaliser une **annotation** ?

Quelques prespectives...

- **Enrichir** les motifs morpho-syntaxiques (syntaxe, coréférence)
- **Modèles de prédiction** qui tirent mieux parti des règles

$\Rightarrow$ Merci pour votre attention !