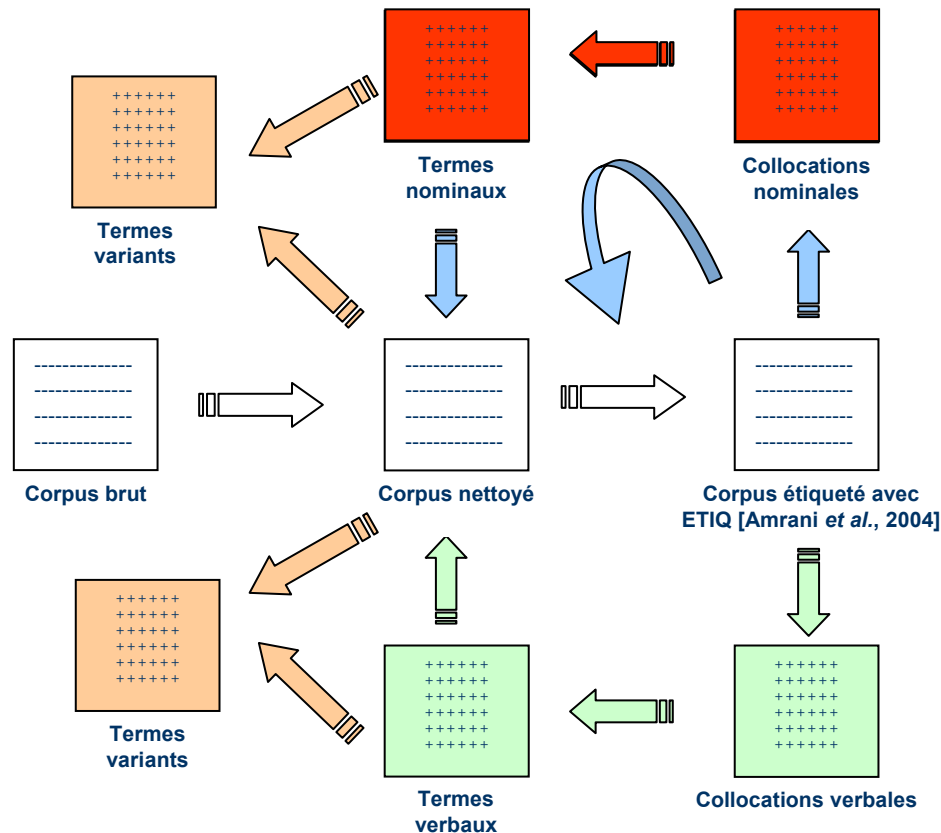


# Étude de Mesures de Qualité pour Classer les Termes Extraits de Corpus Spécialisés

Mathieu Roche,  
Oriane Matte-Tailliez,  
Yves Kodratoff  
**LRI (Orsay)**



# Processus global en terminologie



## Exemples :

Termes *Nom-Prép-Nom* avec l'information mutuelle

1. beurre de karité (3)
2. jéjunum de rat (3)
3. puy en velay (3)
4. chalon sur saône (4)

Termes *Nom-Prép-Nom* avec l'information mutuelle au cube

1. mise en place (111)
2. traitement de texte (57)
3. tableau de bord (23)
4. contrat de qualification (31)

# Corpus

- Corpus de Ressources Humaines (société *PerformanSe*) – 3784 Ko (en français)
- Corpus de CV (Groupe *VediorBis*) – 2470 Ko (en français)
- Corpus d'introductions d'articles sur la Fouille de Données – 369 Ko (en anglais)
- Corpus de résumés d'articles sur la Biologie Moléculaire – 9424 Ko (en anglais)

# Plan de l'exposé

- Présentation de quelques mesures
- Expérimentations
  - Mesure d'évaluation
  - Protocole expérimental
  - Résultats
- Conclusions et perspectives

# Quelques mesures (1/4)

- Information Mutuelle [Church et Hanks, 90]

$$IM(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \quad \Rightarrow \quad I(x, y) = \log_2 \frac{nb(x, y)}{nb(x)nb(y)}$$

- Information Mutuelle au Cube [Daille, 94]

$$I^3(x, y) = \log_2 \frac{nb^3(x, y)}{nb(x)nb(y)}$$

## Quelques mesures (2/4)

- Mesure d'Association [Jacquemin, 97] :
  - isobarycentre des valeurs normalisées de l'information mutuelle et du nombre d'occurrences.

$$a(x, y) = \frac{1}{2} \frac{I(x, y)}{I_M - I_m} + \frac{1}{2} \frac{nb(x, y)}{nb_M - nb_m}$$

$$I_M = \max I(p, q), \quad I_m = \min I(p, q)$$

$$nb_M = \max nb(p, q), \quad nb_m = \min nb(p, q)$$

# Quelques mesures (3/4)

- Coefficient de Dice [Smadja, 96]

$$Dice(x, y) = \frac{2P(x, y)}{P(x) + P(y)}$$

$$\Rightarrow D(x, y) = \frac{2nb(x, y)}{nb\_type(y).nb(x) + nb\_type(x).nb(y)}$$

# Quelques mesures (4/4)

- Rapport de Vraisemblance [Dunning, 93]

	$y$	$y'$ avec $y' \neq y$
$x$	a	b
$x'$ avec $x' \neq x$	c	d

$$\begin{aligned} RV = & a \log(a) + b \log(b) + c \log(c) + d \log(d) - (a+b) \log(a+b) \\ & - (a+c) \log(a+c) - (b+d) \log(b+d) - (c+d) \log(c+d) \\ & + (a+b+c+d) \log(a+b+c+d) \end{aligned}$$

# Plan de l'exposé

- Présentation de quelques mesure
- **Expérimentations**
  - Mesure d'évaluation
  - Protocole expérimental
  - Résultats
- Conclusions et perspectives

# Expérimentations : mesures d'évaluation

- La précision (1)

$$précision = \frac{\text{nombre de collocations extraites pertinentes}}{\text{nombre de collocations extraites}}$$

1. real world  
2. neural network  
3. frequent itemset  
4. remote sensing  
5. naive bayes  
...

*Collocations extraites*



1. **real world**  
2. **neural network**  
3. **frequent itemset**  
4. remote sensing  
5. **naive bayes**  
...

# Expérimentations : mesures d'évaluation

- La précision (2)

Les courbes d'élévation (« lift chart ») : variation de la précision en fonction du nombre de collocations proposées à l'expert.

- Ne connaissant pas l'ensemble des collocations pertinentes, le rappel n'est pas calculé.

# Expérimentations : protocole expérimental

- Corpus de Fouille de Données, de CV, de Ressources Humaines : **termes pertinents** qui sont traces de concepts (*resp. 642, 412 et 2960 termes sur les corpus de Fouille de Données, de CV et des Ressources Humaines*).
- Corpus de Biologie Moléculaire : **termes pertinents et non valides** (*7057 termes*).

# Expérimentations : corpus de Fouille de Données, de CV et des Ressources Humaines

- Elagage à 3

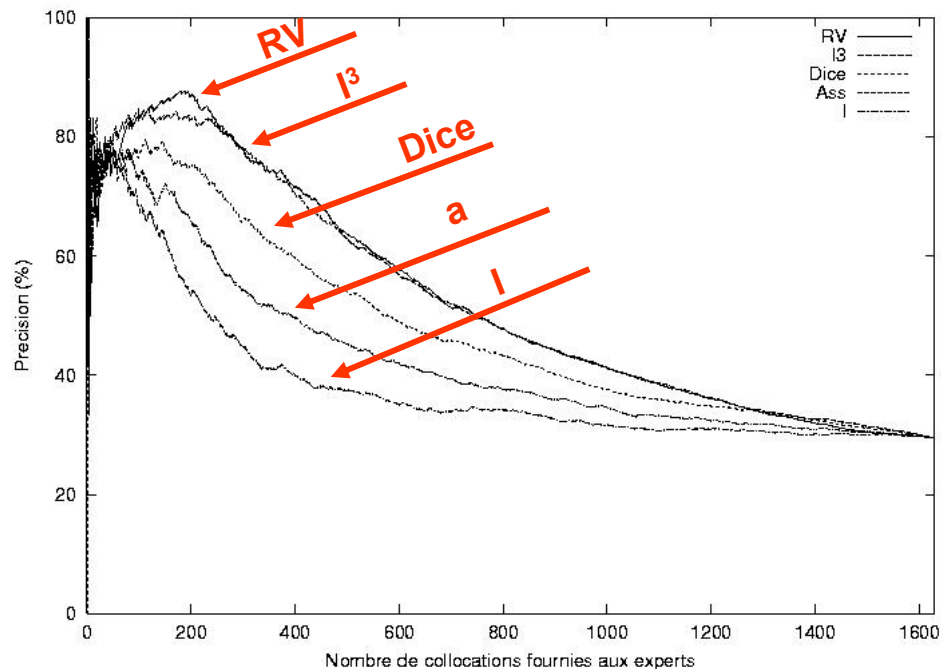
	<i>Nb collocations</i>			<i>Nb collocations après élagage</i>		
	<i>FD</i>	<i>RH</i>	<i>CV</i>	<i>FD</i>	<i>RH</i>	<i>CV</i>
Nom-Prep-Nom	313	4703	3634	7	1268	307
Nom-Nom	2070	98	1781	223	11	162
Adjectif-Nom	2411	1260	1291	176	478	103
Nom-Adjectif	X	5768	3455	X	1628	448

*Exemples :*

emploi solidarité  
action communication  
fichier client  
service achat  
...

# Expérimentations : corpus des Ressources Humaines (*relation Nom-Adjectif*)

- Courbes d'élévation avec cinq mesures.



# Expérimentations : corpus de Biologie Moléculaire

- Elagage à 4.
  - Expérimentations avec la relation Nom-Nom.
  - Les collocations non expertisées ne sont pas prises en compte dans le calcul de la précision.
- ***Résultat similaire aux 3 autres corpus : le Rapport de Vraisemblance est la mesure la plus adaptée pour notre tâche.***

# Paramétrage de la mesure d'association

(relation Nom-Adjectif du corpus des Ressources Humaines)

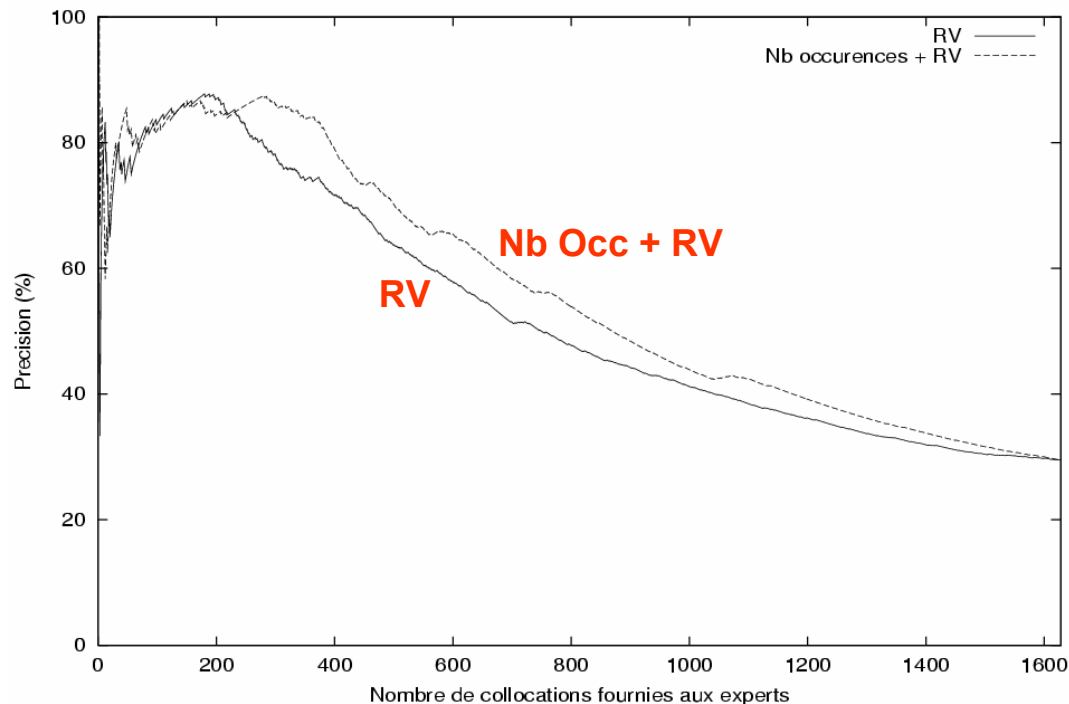
- Ajout d'un paramètre  $\lambda$  à la mesure d'association :

$$a_{\lambda}(x, y) = \lambda \frac{I(x, y)}{I_M - I_m} + (1 - \lambda) \frac{nb(x, y)}{nb_M - nb_m}$$

Collocations extraites	$\lambda = 0$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$
20 %	84.0 %	83.6 %	71.6 %	62.4 %	57.5 %	52.3 %
40 %	61.1 %	57.1 %	50.3 %	45.9 %	42.2 %	40.3 %
60 %	44.7 %	42.4 %	39.5 %	37.3 %	35.6 %	34.7 %
	$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$	$\lambda = 0.9$	$\lambda = 1$	
20 %	49.8 %	47.3 %	45.5 %	43.3 %	42.4 %	
40 %	39.0 %	38.0 %	37.1 %	35.4 %	34.1 %	
60 %	33.8 %	33.2 %	32.5 %	32.0 %	31.6 %	

# Nombre d'occurrences + Rapport de Vraisemblance *(relation Nom-Adjectif du corpus des Ressources Humaines)*

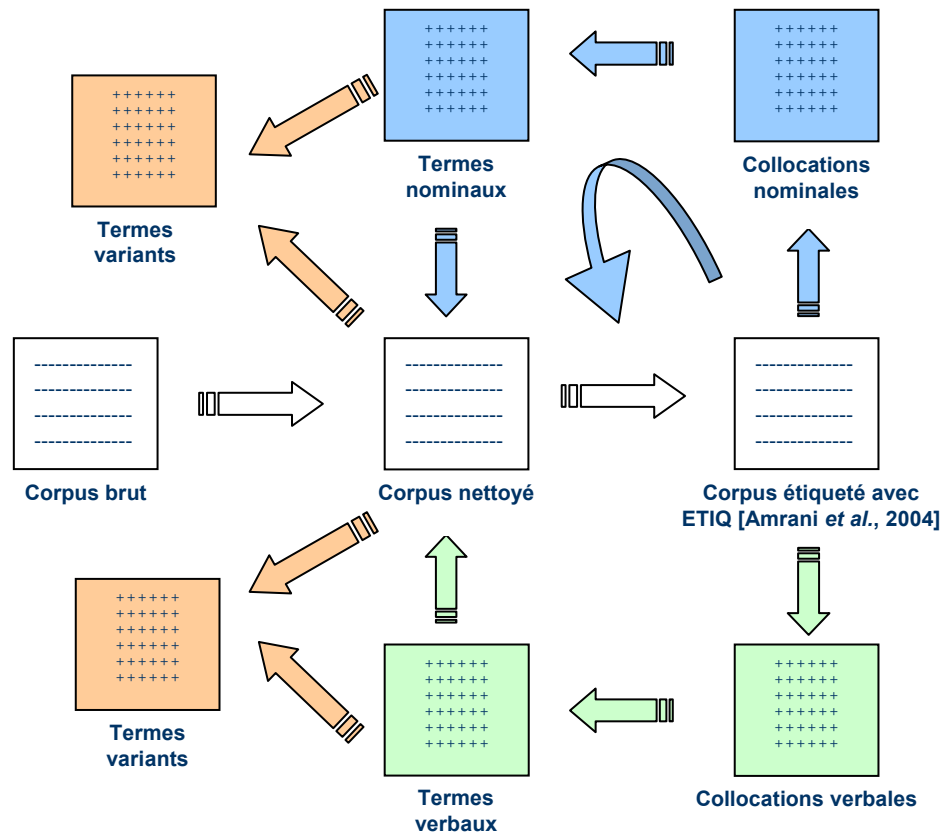
- Classement selon le nombre d'occurrences + une mesure statistique pour les collocations ayant le même nombre d'occurrences.



# Conclusions et perspectives

- Les mesures privilégiant les collocations ayant un nombre d'occurrences important donnent de meilleurs résultats pour notre tâche (extraire des collocations qui sont des traces de concepts).
- L'Information Mutuelle extrait des collocations rares : utilisation de l'Information Mutuelle pour quelle tâche ?
- *Perspectives* : Combinaison de mesures pour améliorer la précision.

# Processus global en terminologie



## Exemples :

Termes *Nom-Prép-Nom* avec l'information mutuelle

1. beurre de karité (3)
2. jéjunum de rat (3)
3. puy en velay (3)
4. chalon sur saône (4)

Termes *Nom-Prép-Nom* avec l'information mutuelle au cube

1. mise en place (111)
2. traitement de texte (57)
3. tableau de bord (23)
4. contrat de qualification (31)