



# Extraction de règles d'association de concepts à partir de différents corpus spécialisés

Yves KODRATOFF <sup>1</sup>, Oriane MATTE-TAILLIEZ <sup>1</sup>, Mathieu ROCHE <sup>1</sup>, Jérôme AZÉ <sup>1</sup>, Ahmed AMRANI <sup>2</sup>

<sup>1</sup> Équipe Inférence et Apprentissage du Laboratoire de Recherche en Informatique - Université Paris-Sud - Orsay – France

<sup>2</sup> ESIEA Recherche, 9 rue Vésale, 75005 Paris

E-mail: {yk,oriane,roche,aze,amrani}@lri.fr

## Première étape : Nettoyage

- Interface pour normaliser des corpus.

## Deuxième étape : Etiquetage

- Interface pour l'ajout de règles afin d'enrichir le lexique de l'étiqueteur de Brill (*Brill, 1994*) et pour l'ajout de nouvelles étiquettes (par exemple, l'étiquette « DNA-motif dégénéré » en Biologie Moléculaire).

## Troisième étape : Extraction de la terminologie

- Six types de candidats termes sont extraits : Adjectif-Nom, Nom-Adjectif, Nom-Nom, Formule-Nom, Nom-Préposition-Nom, Nom-Verbe\_gérondif.
- Les candidats termes sont sélectionnés itérativement par leur taille croissante.

## Quatrième étape : Caractérisation des occurrences de concepts dans les textes

- Les traces des concepts sont des collocations.
- Induction en extension sur les concepts décrits en extension par l'expert.

## Cinquième étape : Extraction de connaissance

- Extraction des règles d'association de concepts en utilisant la mesure de moindre contradiction (Azé & Kodratoff, 2002) et l'intensité d'implication normalisée (Lerman & Azé, 2003).
- Extraction d'information par patrons d'extraction et résolution d'anaphores complexes.



