

Extraction de la Terminologie du Domaine : Etude de Mesures sur un Corpus Spécialisé Issu du Web

Mathieu ROCHE, Oriane MATTE-TAILLIEZ, Jérôme AZÉ, Yves KODRATOFF

Équipe Inférence et Apprentissage du Laboratoire de Recherche en Informatique - Université Paris-Sud - Orsay - FRANCE
E-mail: {roche, oriane, aze, yk}@lri.fr

Le but de nos travaux est d'extraire les termes les plus pertinents d'un domaine (par exemple, à partir d'un corpus issu du Web traitant de Biologie Moléculaire) afin de construire une classification conceptuelle. Une telle tâche s'effectue en quatre étapes.

Première étape : Collecte du corpus

- Collecte du corpus à partir du site « National Center for Biotechnology Information » (<http://www.ncbi.nlm.nih.gov/>).
- Ce corpus représente 6000 résumés d'articles sur la Biologie Moléculaire.

Deuxième étape : Nettoyage

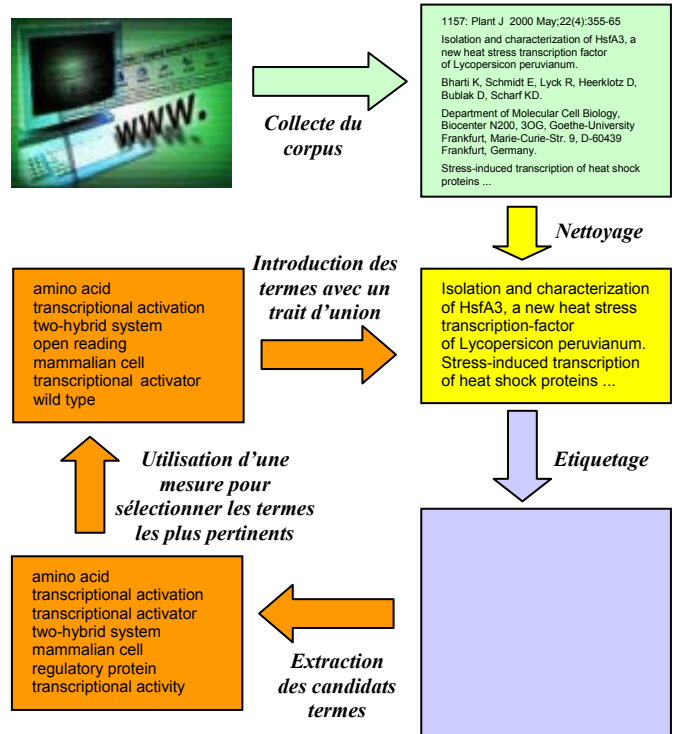
- Interface pour normaliser des textes.

Troisième 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 « Formule » en Biologie Moléculaire).

Quatrième étape : Extraction de la terminologie

- Cinq types de candidats termes sont extraits : Adjectif-Nom, Nom-Nom, Formule-Nom, Nom-Préposition-Nom, Nom-Verbe_gérondif.
- Les candidats termes les plus pertinents sont retenus itérativement selon une mesure.



Extraction de la terminologie avec différentes mesures

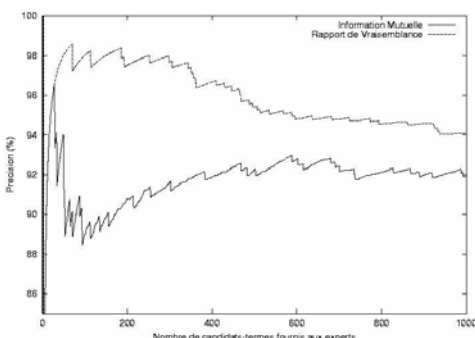
Deux exemples de mesure

Termes binaires formés avec les mots « x » et « y ».

- Information mutuelle :** $IM = \log_2(P(x,y)/P(x)P(y))$
- Rapport de vraisemblance :**

	y	y' avec y' ≠ y
x	a	b
x' avec x' ≠ x	c	d

$$RV = a \log(a) + b \log(b) + c \log(c) + d \log(d) - (a+b) \log(a+b) - (a+c) \log(a+c) - (b+d) \log(b+d) - (c+d) \log(c+d) + (a+b+c+d) \log(a+b+c+d)$$



Ensemble des mesures testées

Nb de termes proposés à l'expert	100	200	500	1000
Information mutuelle	89.0 (100%)	90.8 (76%)	92.2 (51%)	91.9 (43%)
Information mutuelle au cube	96.0 (100%)	97.5 (100%)	94.0 (87%)	94.1 (61%)
Mesure d'association	90.0 (100%)	91.2 (80%)	93.0 (53%)	92.5 (46%)
Coefficient de Dice	92.0 (100%)	92.9 (92%)	92.6 (73%)	93.0 (53%)
Rapport de vraisemblance	98.0 (100%)	97.5 (100%)	95.4 (92%)	94.1 (62%)
J-mesure	89.0 (27%)	89.1 (23%)	89.4 (26%)	95.2 (42%)
Conviction	96.9 (97%)	97.4 (79%)	97.2 (57%)	95.2 (42%)
Sebag-Schoenauer	93.1 (58%)	94.9 (60%)	94.7 (53%)	94.7 (43%)
Moindre contradiction	96.0 (99%)	96.1 (77%)	95.3 (43%)	95.9 (32%)
Intensité d'implication	99.0 (100%)	96.6 (89%)	95.5 (67%)	93.0 (50%)
Intensité d'implication nor.	99.0 (100%)	96.6 (89%)	95.5 (67%)	92.8 (50%)

Précision (en %) selon le nombre de termes extraits pour la relation adjectif-nom. Le pourcentage entre parenthèses représente le pourcentage de termes analysés par l'expert parmi les termes extraits. Pour chaque nombre de termes proposés à l'expert, nous notons « en gras » les 4 mesures ayant la précision la plus élevée.

Laboratoire de
Recherche en
Informatique

