



HAL
open science

Extraction automatique d'affixes pour la reconnaissance d'entités nommées chimiques

Yoann Dupont, Isabelle Tellier, Christian Lautier, Marco Dinarelli

► **To cite this version:**

Yoann Dupont, Isabelle Tellier, Christian Lautier, Marco Dinarelli. Extraction automatique d'affixes pour la reconnaissance d'entités nommées chimiques. EGC, Jan 2016, Reims, France. hal-01476792

HAL Id: hal-01476792

<https://hal.science/hal-01476792>

Submitted on 25 Feb 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Extraction automatique d'affixes

pour la reconnaissance d'entités nommées chimiques



Yoann Dupont^{1,2}, Isabelle Tellier¹, Christian Lautier², and Marco Dinarelli¹

¹Laboratoire Lattice, UMR 8094, 1 rue Maurice Arnoux, 92120 Montrouge

²Temis SA, 207 rue de Bercy, 75012 Paris



I. Résumé

Nous détaillons ici une approche permettant de détecter des affixes à partir de dictionnaires en se basant sur l'algorithme de la plus longue sous-chaine commune, dans le cadre de la reconnaissance d'entités nommées chimiques sur CHEMDNER (Krallinger et al., 2015). Nous verrons ensuite des méthodes de sélection et de tri pour choisir les plus pertinents.

II. Corpus CHEMDNER

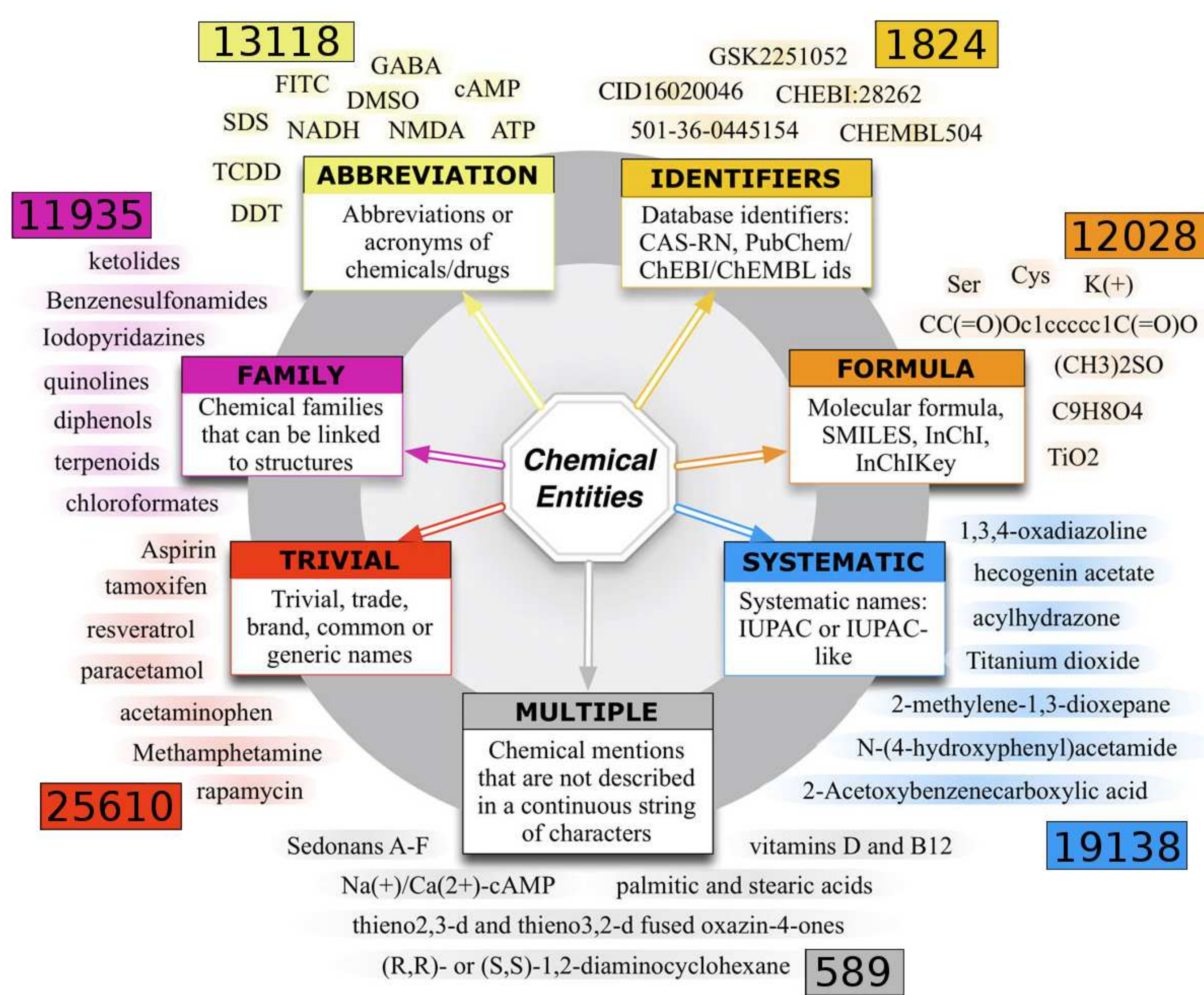


Figure 1: Les entités du corpus CHEMDNER ainsi que leurs comptes

III. Annotation de séquences avec des CRF

étiquette	B-TRIVIAL	I-TRIVIAL	O	O	O	B-FORMULA	B-FAMILY	O	[...]
tokens	Glucogenin	E	,	a	new	C21	steroid	from	[...]
préfixe 3	Gla	#	#	#	new	C21	ste	fro	[...]
suffixe 3	nin	#	#	#	new	C21	oid	rom	[...]

Figure 2: CRF → système à base de traits pondérés pour retrouver la séquence "étiquette"

$$f(\text{B-TRIVIAL} | \text{mot}=\text{Glucogenin}) = \text{poids}_1 \quad f(\text{I-TRIVIAL} | \text{mot}=\text{Glucogenin}) = \text{poids}_2$$
$$f(\text{B-FAMILY} | \text{préfixe } |3| = \text{ste}) = \text{poids}_x \quad f(\text{I-FAMILY} | \text{préfixe } |3| = \text{ste}) = \text{poids}_{x+1}$$

$$\text{etiquette}(\text{traits}) = \operatorname{argmax}_{E \in \text{Etiquettes}} \sum_{t \in \text{traits}} f(E|t)$$

Figure 3: vision simplifiée du fonctionnement d'un CRF

IV. Extraction d'affixes

	m	e	n	t	h	o	n	e
o	0	0	0	0	0	0	0	0
r	0	0	0	0	0	0	0	0
o	0	0	0	0	0	0	1	0
t	0	0	0	0	1	0	0	0
e	0	0	1	0	0	0	0	0
n	0	0	0	2	0	0	0	1
o	0	0	0	0	0	0	1	0
n	0	0	0	0	1	0	0	0
e	0	0	1	0	0	0	0	3

Table 1: matrice de calcul de sous-séquences

préfixes	infixes	suffixes
∅	{e,o,n,t,en}	{one}

Table 2: affixes extraits

- liste d'affixes candidats
- liste bruitée : besoin d'élaguer !

V. Sélection et tri des affixes

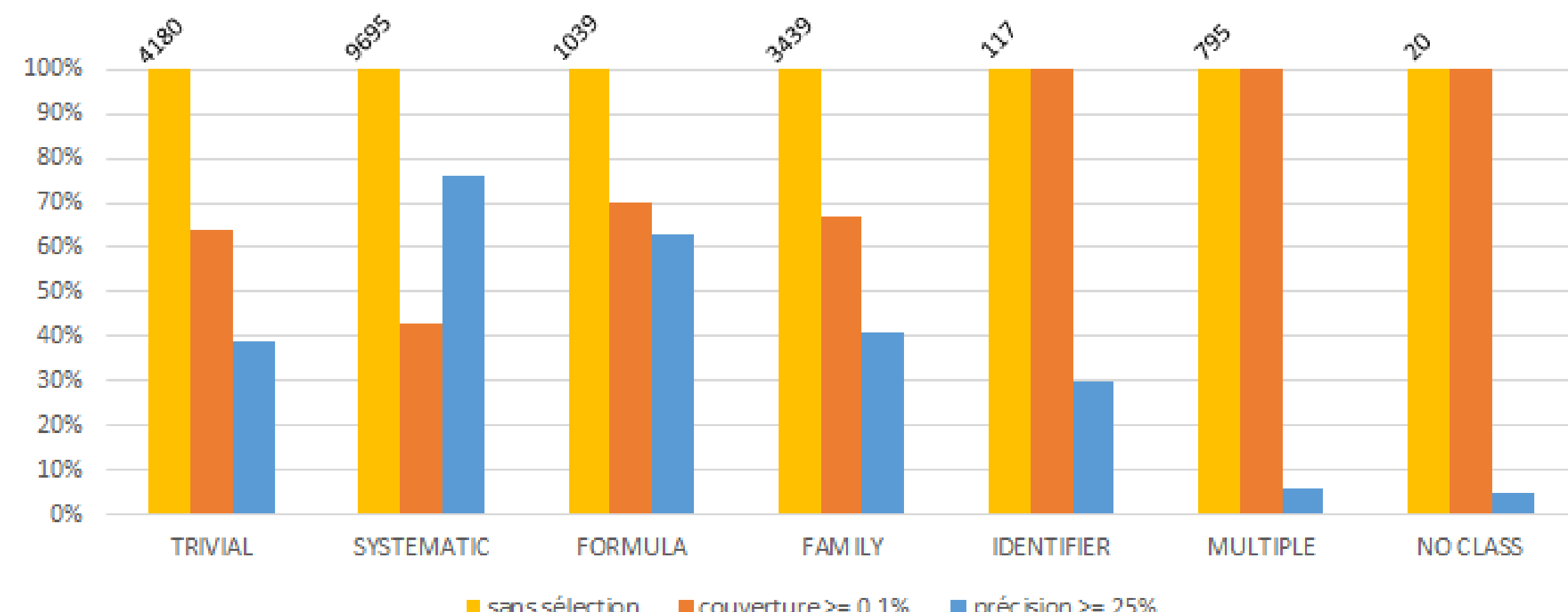


Figure 4: Réduction du volume des affixes selon le score calculé sur le corpus d'entraînement

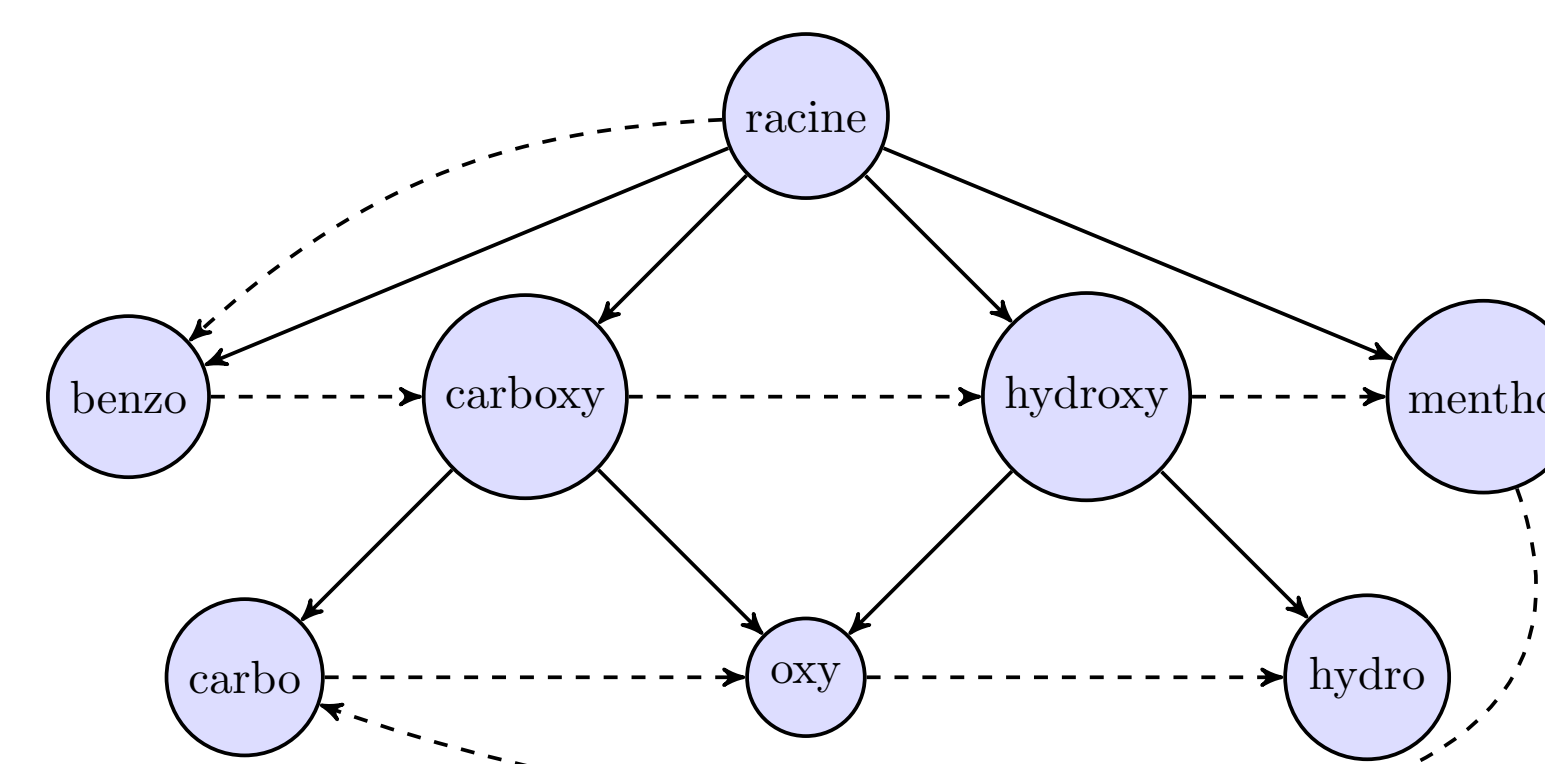
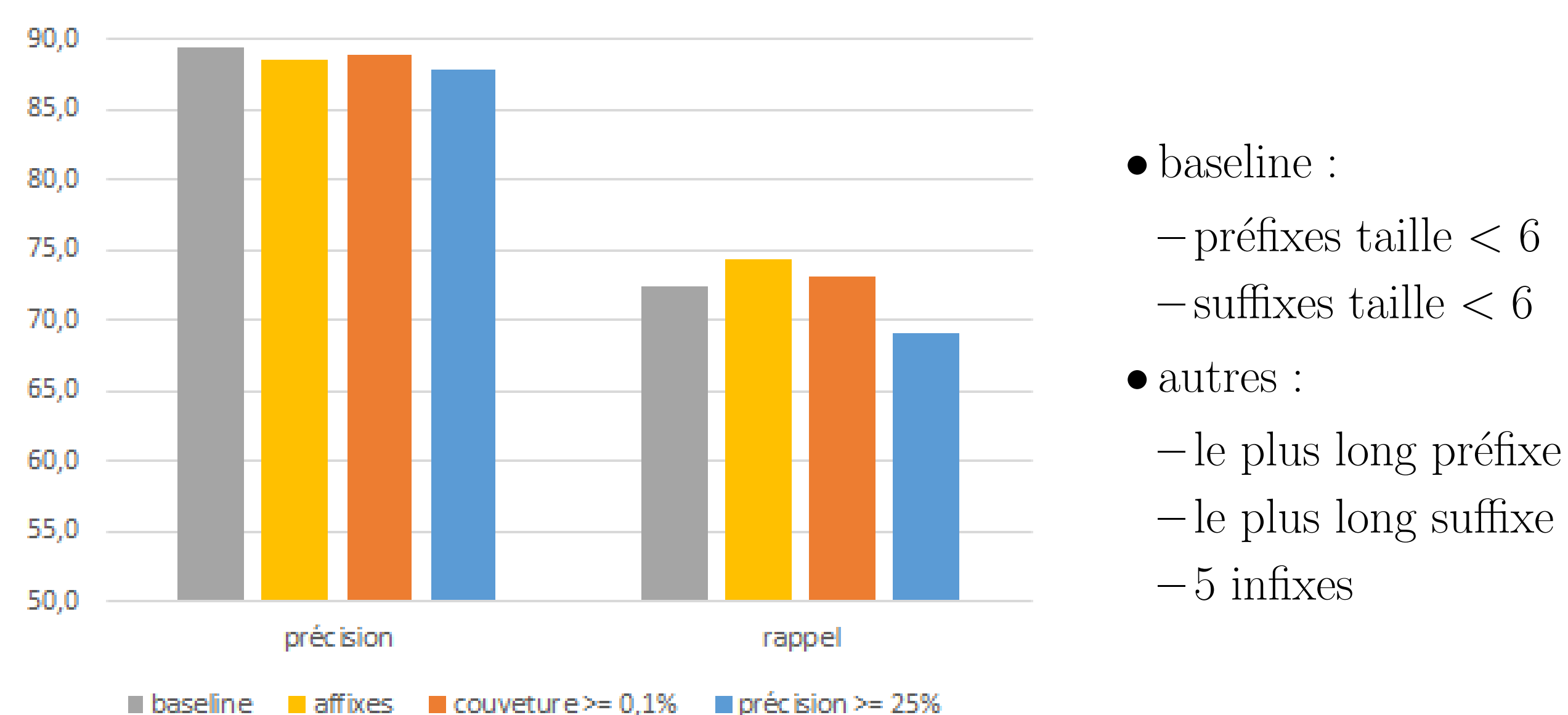


Figure 5: Le treillis des affixes. Représente la relation "X est une sous-chaine stricte de Y". Les affixes sont sélectionnés selon une recherche en largeur (pointillés)

VI. Résultats



VII. Conclusions et perspectives

- approche pour trouver automatiquement des affixes
- entités extraites du corpus d'entraînement → utiliser PubChem (Wang et al., 2009)
- sélection : probabilité conditionnelle entre deux affixes apparentés (Zhang and Lee, 2006)
- intégrer affixes dans des méthodes vectorielles (SVM, RN)
- appliquer la méthode pour trouver automatiquement des termes déclencheurs

IUT de Reims Châlon Charleville
IUT de de Troyes
Université de technologies de Troyes
Université d'Angers

Références

Krallinger, M., O. Rabal, F. Leitner, M. Vazquez, D. Salgado, Z. Lu, R. Leaman, Y. Lu, D. Ji, D. M. Lowe, et al. (2015). **The CHEMDNER corpus of chemicals and drugs and its annotation principles.** *Journal of cheminformatics* 7(Suppl 1), S2.

Wang, Y., J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant (2009). **PubChem: a public information system for analyzing bioactivities of small molecules.** *Nucleic acids research* 37(suppl 2), W623–W633.

Zhang, D. and W. S. Lee (2006). **Extracting Key-substring-group Features for Text Classification.** In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06*, New York, NY, USA, pp. 474–483. ACM.