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

Yoann Dupont[*,**], Isabelle Tellier[*], Christian Lautier[**], Marco Dinarelli[*]

[*]Lattice - UMR 8094, 1 rue Maurice Arnoux, 92120 Montrouge
yoann.dupont@etud.sorbonne-nouvelle.fr
isabelle.tellier@univ-paris3.fr
marco.dinarelli@ens.fr
[**]Temis SA, 207 rue de Bercy, 75012 Paris
yoann.dupont,christian.lautier@temis.com

Résumé. Nous détaillerons ici une approche permettant de détecter des affixes à partir de dictionnaires en se basant sur l'algorithme de la plus longue sous-chaîne commune, dans le cadre de la reconnaissance d'entités nommées chimiques sur CHEMDNER. Nous verrons ensuite des méthodes de sélection et de tri afin de les intégrer au mieux dans un système d'apprentissage automatique.

1 Introduction

Nous nous sommes inétressés à la tâche CEM (Chemical Entity Mention recognition) du corpus CHEMDNER (Krallinger et al., 2015). CEM décrit huit entités distinctes : les noms de marque ou génériques (TRIVIAL), les noms complets (SYSTEMATIC), les abbréviations (ABBREVIATION), les formules (FORMULA), les familles d'entités (FAMILY), les identifiants (IDENTIFIER), les groupes d'entités (MULTIPLE) et les entités dont la classe n'a pas pu être déterminée (NO_CLASS). Pour ce faire, nous avons utilisé un CRF enrichi avec des affixes détectés automatiquement puis pondérés et ordonnés. Notre méthode est asez proche de celle de Zhang et Lee (2006), où des sous-chaînes sont extraites automatiquement pour servir de features à un système discriminant (SVM) pour la classification de documents.

2 Extraction d'affixes

Pour chaque type d'entité, nous extrayons du corpus d'entraînement l'ensemble de ses instances et appliquons, sur chaque couple, l'algorithme de la plus longue sous-chaîne commune. Cet algorithme crée une matrice où l'ensemble des sous-chaînes communes sont reconnues. Il est possible de diviser ces affixes en trois catégories : préfixes, suffixes et infixes. Nous avons pondéré nos traits en utilisant leur précision et couverture. La précision d'un trait est la proportion de mots reconnus dans la bonne classe par rapport au nombre de mots reconnus. La couverture celle des mots reconnus parmi l'ensemble des mots d'une classe donnée. Nous avons aussi créé une structure afin de classer les infixes. Il s'agit d'un graphe orienté acyclique