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

(DAG) construit selon la relation d'ordre « X est une sous-chaîne stricte de Y », illustré dans la figure 1.

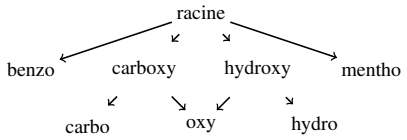

FIG. 1: exemple de traits hiérarchisés

3 Résultats

Notre baseline utilise des préfixes et suffixes de taille 1 à 5. Les score que nous avons obtenus sont les suivants : 89,40% de précision, 72,41% de rappel et 80,01% de f-mesure. En comparaison, nous avons utilisé uniquement les plus longs préfixe et suffixe. Nous avons aussi utilisé un ensemble de 5 infixes triés selon la figure 1. Nous avons effectué trois expériences : (a) sans sélection, (b) sélection par précision, (c) sélection par couverture. (a) est l'expérience qui a donné le meilleurs résultats : 88,48% de précision, 74,37% de rappel et 80,82% de f-mesure. Suivie de (c) : 88,82% de précision, 73,15% de rappel et 80,23% de f-mesure. Finalement, (b) : 87,75% de précision, 69,03% de rappel et 77,28% de f-mesure, qui est moins bonne que la baseline. Les présélections n'ont pas apporté d'amélioration globale par rapport à l'ajout de l'intégralité des traits. Pour l'expérience (a), L'entité ayant eu la meilleure amélioration globale est SYSTEMATIC (+1.92), d'abord sur les entités connues (+2.18) puis les inconnues (+1.55). La meilleure amélioration sur les entités inconnues s'est faite sur FORMULA (+2.05). Les plus grosses pertes sur les entités inconnues sont sur ABBREVIATION (-7.75), IDENTIFIER (-4.8) et TRIVIAL (-3.86). Malgré une amélioration globale, nous voyons que notre approche peut encore être améliorée.

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.

Zhang, D. et 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.

Summary

In this article we explain an automatic approach to detect affixes from entries in a dictionary using the longest common substring algorithm, in the context of chemical named entity recognition on the CHEMDNER corpus. We then show selection and sorting methods in order to better integrate them in a machine learning system.