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. Dans cet article nous détaillons une approche permettant de détecter des affixes et des termes déclencheurs à partir de dictionnaires de façon automatique 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.

2 Extraction d'affixes

Pour chaque type d'entité, nous extrayons du corpus d'entraînement l'ensemble de ses instances et appliquons un algorithme sur l'ensemble des couples d'entrées pour extraire un ensemble d'affixes candidats. Ces ensembles peuvent être dinstigués en trois catégories : les préfixes, les suffixes et les infixes. L'algorithme que nous avons utilisé comme base extraire les affixes est celui de la plus longue sous-chaîne commune [1]. Nous avons utilisé la matrice générée pour récupérer l'ensemble des sous-chaînes communes pour chaque couple. Nous avons pondéré nos traits en utilisant des scores de précision et de couverture. La précision d'un trait se définit alors comme la proportion de tokens reconnus dans la bonne classe par rapport

1. https://en.wikipedia.org/wiki/Longest_common_substring_problem

au nombre de tokens reconnus, et la couverture comme la proportion de tokens reconnus parmi l'ensemble des tokens d'une classe donnée. Nous avons également créé une structure afin de classer les affixes. Il s'agit d'un graphe orienté acyclique (DAG) construit selon la relation d'ordre « X est une sous-chaîne stricte de Y », un tel arbre est illustré dans la figure 1.

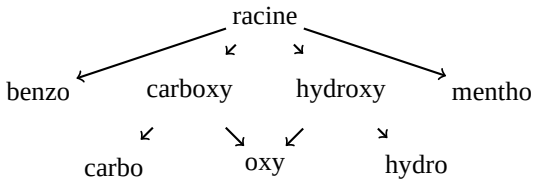

FIG. 1 – *exemple de traits hiérarchisés*

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. Bien que les affixes soient générés éntité par entité, nous avons regroupé tous les ensembles pour ajouter de l'information au CRF, ce dernier établissant lui-même les correspondances trait/entité. L'utilisation des affixes a amélioré notre baseline. 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 se font 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.

Summary

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