Annotation d'Entités Nommées par Extraction de Règles de Transduction

Damien Nouvel *,** Arnaud Soulet *,***

*Université François Rabelais Tours, Laboratoire d'Informatique
** damien.nouvel@univ-tours.fr,
** arnaud.soulet@univ-tours.fr,

Résumé. La reconnaissance d'entités nommées est une problématique majoritairement traitée par des modèles spécifiés à l'aide de règles ou par apprentissage numérique. Les premiers ont le désavantage d'être coûteux à développer pour obtenir une couverture satisfaisante, les seconds sont souvent difficiles à interpréter par des experts (linguistes). Dans cet article, nous présentons une approche, dont l'objectif est d'extraire des règles symboliques discriminantes qu'un humain puisse consulter. A partir d'un corpus de référence, nous extrayons des règles de transduction, dont seules les plus informatives sont retenues. Elles sont ensuite appliquées pour effectuer une annotation : à cet effet, un algorithme recherche parmi les annotations possibles celles de meilleure qualité en termes de couverture et de probabilité. Nous présentons les résultats expérimentaux et discutons de l'intérêt et des perspectives de notre approche.

1 Introduction

Parmi les tâches d'extraction d'information, la Reconnaissance d'Entités Nommées (REN) consiste à reconnaître (rechercher et catégoriser) toutes les Entités Nommées (EN) d'un texte : les expressions *univoques* et *référentiellement autonomes* (Ehrmann, 2008). Par simplification, ces unités correspondent intuitivement aux noms propres : personnes, lieux et organisations. En pratique seront aussi recherchées les expressions numériques et les expressions de temps, considérées comme "descriptions définies".

La REN est une tâche étudiée depuis une quinzaine d'années. Initialement symboliques, les approches sont aujourd'hui majoritairement tournées vers des modèles numériques par recherche de traits discriminants à l'aide d'apprentissage automatique (e.g., Chaînes de Markov, CRF, SVM). Cependant, ces méthodes permettent difficilement de capitaliser la connaissance modélisée.

Malgré ces récents développements, les dernières campagnes d'évaluation en français montrent que les systèmes symboliques demeurent plus performants, à condition d'utiliser des bases de connaissances suffisamment riches. Ces dernières sont généralement constituées de lexiques (recensant les formes connues de noms propres) et de règles pour reconnaître des entités nommées (spécifiées par des grammaires). Entre autres, les transducteurs sont des expressions régulières permettant d'insérer dans un