Étude comparative pour l'analyse de requêtes complexes dans le domaine du pneumatique

Abdenacer Keraghel**, Khalid Benabdeslem*
Bruno Canitia**

*Université Lyon 1, 43 Boulevard du 11 novembre 1918, Villeurbanne cedex 69622
khalid.benabdeslem@univ-lyon1.fr,
**Lizeo IT, 42 Quai Rambaud, 69002 Lyon
abdenacer.keraghel, bruno.canitia@lizeo-group.com
https://www.lizeo-group.com

Résumé. La recherche et l'extraction d'information dans une séquence textuelle (article scientifique, requête sur un moteur de recherche, post sur un forum de discussion) nécessitent un processus de reconnaissance d'entité nommée (REN). Cependant, les données disponibles pour effectuer ce processus varient selon leur nature et le domaine d'étude. Dans cet article, nous nous intéressons d'une part aux performances des systèmes de reconnaissance d'entité nommée, et d'autre part à leur complexité ainsi qu'à leur capacité à traiter des données de différentes origines. Une étude comparative entre plusieurs approches issues de l'état de l'art, appliquée à différents types de données (requêtes d'un moteur de recherche et posts de forums de discussion) liées au domaine du pneumatique, est proposée afin de sélectionner l'approche qui s'adapte le mieux à notre cas d'usage. Pour cela, nous nous appuierons sur les résultats de métriques d'évaluation des modèles d'apprentissage automatique telles que la précision, le rappel et la F-mesure.

1 Introduction

Une entité nommée est une unité linguistique référentielle, associée à des objets tels que : les noms de personnes (PER) comme « Barack Obama », les organisations (ORG) comme « Google », les lieux (LOC) comme « Paris », et autres (MISC) comme « Championship ». La reconnaissance d'entité nommée est une sous-tâche de l'activité d'extraction d'information qui consiste à identifier ces objets dans une séquence textuelle. Dans cet article, nous nous intéressons au domaine du pneumatique, dont les objets sont désignés par la marque ou le type de véhicule, la dimension ou la saison du pneu, etc.

Un système de reconnaissance d'entité nommée sans ambiguïté a pour but d'analyser (via une tokenisation [1] et un calcul de score de confiance) les données en entrée afin de détecter leurs classes associées. En l'espèce, un token [2] d'une séquence peut appartenir à plusieurs classes,

1. L'opération consiste à découper un texte en token, le plus souvent des mots.
2. C'est une unité lexicale.