Reconnaissance de concepts basée sur l'apprentissage

Wahiba Ben Abdessalem Karâa *, Bilel Bouchamia**

*, ** ISG Tunis 41 avenue de la Liberté Cité Bouchoucha, Le Bardo 2000 , Tunis
*Wahiba.abdessalem@isg.rnu.tn
** bilbouch7@hotmail.com

1 Introduction

Le papier présente une approche permettant de reconnaître des concepts (entités nommées) qui s'appuie sur l'identification des contextes à gauche ou à droite d'un concept. Notre approche permet d'identifier et de classifier les contextes en se basant sur une phase d'apprentissage qui exploite des documents issus du web.

Le papier est organisé de la manière suivante : dans le paragraphe 2, nous décrivons notre démarche. Dans le paragraphe 3 nous présentons quelques résultats des tests de validation de notre approche.

2 Description de la demarche

2.1 Construction de corpus d'apprentissage

La première étape de notre démarche consiste à construire un corpus de documents, appelé corpus d'apprentissage. Pour ce faire, l'utilisateur doit fournir une requête comportant un ensemble d'instances du concept (exemples d'apprentissage). Pour le concept "Capitale" par exemple, il faut fournir, des noms de capitales tels que *"Paris", "Rome", "Tunis", "Madrid"*, etc. La requête est destinée aux moteurs de recherche (Yahoo, Google, etc.) en utilisant une API spécifique au moteur. Le moteur renvoie les résultats sous formes de liens vers des pages web. Ces liens sont utilisés par la suite pour récupérer les documents.

2.2 Extraction des contextes

Un contexte est un ensemble de mots qui précédent ou qui suivent les instances d'un concept. L'objectif de ce travail consiste à identifier, dans le corpus d'apprentissage, les contextes les plus pertinents pour identifier un concept donné. Pour ce faire nous avons calculé des scores inspirés de la représentation *TF-IDF* (Salton et al, 1975).

$$SOCC = \frac{OCC}{OCCEXP}, \qquad SSEED = \frac{SEED}{EXP}, \qquad SDOC = \frac{DOCDIF}{DOC}$$

OCC est le nombre d'occurrences d'un contexte dans le corpus. *OCCEXP* est le nombre d'occurrences des exemples d'apprentissage dans le corpus.

SEED, représente le nombre d'instances du concept qui se trouvent avec le contexte. *EXP* représente le nombre d'exemples d'apprentissage utilisés.

DOC désigne le nombre de documents dans le corpus contenant au moins une occurrence