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

du contexte. *DOCDIF* représente le nombre de documents parmi *DOC* qui proviennent de sources différentes. En effet, deux liens URL peuvent référencer un même document.

*SCORE 1 = SOCC*SSEED*SDOC*, est un score qui permet de classifier les contextes par rapport aux autres contextes *(SOCC),* par rapport aux instances du concept *(SSEED)* et par rapport aux documents dans les quels se trouve le contexte *(SDOC).*

Nous avons considéré l'hypothèse qu'un contexte est pertinent s'il apparaît plus souvent avec les exemples d'apprentissage et moins souvent avec d'autres entités textuelles. Pour cette raison, nous avons calculé *SCORE2*, qui représente le poids du contexte par rapport aux autres entités textuelles. $SCORE2 = \frac{OCC}{OCCTOT}$

OCC représente le nombre d'occurrences du contexte avec les exemples d'apprentissage. *OCCTOT* représente le nombre d'occurrences total du contexte dans le corpus.

Nous avons calculé un nouveau score *S1S2 = SCORE1*SCORE2,* qui peut être un nouveau paramètre pour déterminer la pertinence d'un contexte dans le corpus.

3 Test et validation

Nous avons réalisé plusieurs tests avec des concepts différents. Par exemple, pour le concept *"CAPITAL"*, nous avons extrait un corpus de documents à partir de 65 URL, nous avons obtenu 2398 contextes de deux mots à gauche. Les instances sont rencontrées 4264 fois dans le corpus. Le tableau suivant (TAB 1) représente un extrait des résultats :

Contexte	SOCC	SDOC	SSEED	SCORE1	SCORE2	S1*S2
Hotels in	0,0039869	0,4444445	0,5384616	0,00095413	8,5	0,008110106
Map of	0,0028143	0,6	0,5384616	0,000909235	6	0,005455413
Places in	0,0011726	0,75	0,3076923	0,0002706	1	0,0002706
hotels in	0,0021107	0,5714286	0,6153847	0,000742224	0,2093	0,000155348
travel to	0,000469	1	0,1538462	7,21539E-05	2	0,000144308
map of	0,0007036	1	0,2307692	0,000162369	0,12	1,94843E-05

TAB. 1 – Context classification

Références

Salton, G., Wong, A., Yang, C. S.(1975). A vector space model of information retrieval, Proceedings ACM SIGIR conference, v18, n11, pp. 613–620.

Summary

This work concerns concept (named entity) recognition based on score calculation. The purpose is to identify and classify the contexts that are most pertinent to recognize a concept. We apply a learning approach, using training corpus constructed with web documents.