Fonctions de Rang et Fouille du Web pour l'identification et la catégorisation d'Entités Nommées

Mathieu Roche

Fonctions de Rang et Fouille du Web pour l'identification et la catégorisation d'Entités Nommées

Mathieu Roche

LIRMM, CNRS, Université Montpellier 2
161 rue Ada, 34095 Montpellier Cedex 5 - France

Abstract

This paper describes Web-Mining methods in order to identify and classify Named Entities. The proposed methods are based on statistical measures using syntactic and/or semantic knowledge. The combination of these approaches is particularly relevant. The experiments of our methods, based on the study of nearly 500 named entities, required the implementation of approximately 2800 queries from a search engine. The results show the positive aspects of using Web-Mining methods for this kind of task and open up very encouraging future work.

Résumé

Cet article décrit des méthodes de Fouille du Web afin d'identifier et de catégoriser les Entités Nommées. Les méthodes proposées s'appuient sur des mesures statistiques fondées sur des connaissances syntaxiques et/ou sémantiques. La combinaison de ces approches se révèle particulièrement pertinente. Les expérimentations de nos méthodes, qui s'appuient sur l'étude de près de 500 entités nommées, ont nécessité l'exécution de près de 2800 requêtes à partir d'un moteur de recherche. Les résultats montrent l'intérêt d'utiliser des méthodes de Fouille du Web pour ce type de tâche et ouvrent des perspectives tout à fait encourageantes.

Mots-clés : fouille du web, entités nommées, fouille de textes, fonction de rang

1. Introduction

Les Entités Nommées (EN) sont classiquement définies comme les noms de Personnes, Lieux et Organisations. Initialement, une telle définition est issue des campagnes d'évaluation américaines MUC (Message Understanding Conferences) qui furent organisées dans les années 90. Cette série de campagnes consistait à extraire des informations telles que les EN dans différents documents (messages de la marine américaine, récits d'attentats terroristes, etc). Aujourd'hui, de telles campagnes d'évaluation couvrent des tâches très variées sur la base de textes de différents domaines (textes spécialisés en biologie, dépêches d'actualités, blogs, etc). Nous pouvons, entre autres, citer les challenges TREC - Text REtrieval Conference (international) et DEFT - DEfi Fouille de Textes (francophone) qui sont aujourd'hui très actifs dans la communauté «fouille de textes».

Comme le précisent (Daille *et al.*, 2000), les classes de base d'EN définies dans le cadre de MUC peuvent être enrichies. Par exemple, outre les classes relatives aux Personnes, Lieux et

Organisations, (Paik *et al.*, 1994) définissent de nouvelles classes comme Document (logiciels, matériels, machines) et Scientifique (maladie, médicaments, etc).

De nombreuses méthodes permettent d'identifier les EN (Nadeau et Sekine, 2007). Par exemple, des méthodes de fouille de données fondées sur l'extraction de motifs permettent de déterminer des règles (appelées *règles de transduction*) afin de repérer les EN (Nouvel et Soulet, 2011). Ce type de règles utilise des informations syntaxiques propres aux phrases (Brun et Hagège, 2004 ; Nouvel et Soulet, 2011). Par ailleurs, pour identifier les EN, de nombreux systèmes s'appuient sur la présence de majuscules (Daille *et al.*, 2000). Cependant ceci peut se révéler peu efficace dans le cas d'EN non capitalisées et pour le traitement de textes non normalisés (mails, blogs, textes ou fragments de textes inégalement en majuscule ou minuscule, etc).

Des approches récentes s'appuient sur le Web pour établir des liens entre des entités et leur type (ou catégorie). Par exemple, l'approche de (Bonnefoy *et al.*, 2011) repose sur le principe que les distributions de probabilités d'apparition des mots dans les pages associées à une entité donnée sont proches des distributions relatives aux types. Ce même principe est utilisé pour le traitement des EN à partir de données textuelles complexes comme les tweets (Ritter *et al.*, 2011).

Pour l'identification des EN et/ou leur catégorie, de nombreuses approches s'appuient sur des méthodes d'apprentissage supervisé (Tjong Kim Sang et De Meulder, 2003). Ces méthodes d'apprentissage comme les SVM ou les arbres de décision sont souvent utilisées dans le challenge *Conference on Natural Language Learning (CoNLL)*. Les algorithmes exploitent divers descripteurs ainsi que des données expertisées/étiquetées. Les types de descripteurs utilisés sont par exemple les positions des candidats, les étiquettes grammaticales, les informations lexicales (par exemple, majuscules/minuscules), les affixes, l'ensemble des mots dans une fenêtre autour du candidat, etc. (Carreras *et al.*, 2003). Dans l'approche proposée dans cet article et contrairement aux travaux issus de la conférence CoNLL, nous n'utiliserons pas de méthodes d'apprentissage supervisé qui nécessitent des données étiquetées nombreuses. En effet, elles peuvent se révéler difficiles à obtenir, en particulier pour le traitement de données spécialisées.

Dans les travaux présentés dans cet article, notre objectif est de déterminer si un candidat issu d'une liste de termes est une EN. Pour ce faire, nous utiliserons les critères d'unicité référentielle (c'est-à-dire, un nom propre renvoie à une entité référentielle unique) et de stabilité dénominative (c'est-à-dire, peu de variations possibles) établis par (Fort *et al.*, 2009). Par la suite, cet article se focalisera en grande partie sur l'identification de la classe d'EN associée. Les approches décrites utilisent des méthodes de Fouille du Web qui exploitent des informations sémantiques et syntaxiques.

2. Fonction de Rang et Fouille du Web

Les mesures que nous proposons de type Fouille du Web (Web-Mining) pour déterminer des EN et leur catégorie s'appuient sur des informations exogènes qui sont de deux ordres.

Dans un premier temps, nous mesurons la **présence d'associations** de mots sur le Web. Par exemple, les requêtes (en utilisant les guillemets pour effectuer une recherche exacte) : «LIRMM Montpellier» et «LIMSI Montpellier» via le moteur de recherche Google retournent

des résultats dénotant la validité des associations entre les mots *LIRMM* (resp. *LIMSI*) et *Montpellier*. A contrario la requête «LABRI Montpellier» ne retourne aucun résultat montrant que l'association des deux mots composant cette requête est en fait non pertinente.

Cependant, cette seule information de présence/absence n'est pas toujours suffisante car la quantité de pages Web indexées est extrêmement importante (le nombre de pages web indexées par Google est estimé à vingt mille milliards de documents) et engendre, inévitablement, la présence de documents bruités. Ainsi, nous devons aussi prendre en compte **l'intensité de cette association** qui repose sur le nombre de pages retournées par les moteurs de recherche avec les associations recherchées.

Par exemple, les mêmes requêtes énoncées précédemment («LIRMM Montpellier» et «LIMSI Montpellier») retournent un nombre de pages différent (respectivement 11400 et 1) montrant une intensité de l'association qui n'est clairement pas du même ordre.

Dans ce même cadre, l'algorithme PMI-IR (Pointwise Mutual Information and Information Retrieval) de (Turney, 2001) consiste à interroger le Web via le moteur de recherche AltaVista pour déterminer des synonymes appropriés. A partir d'un terme donné noté *mot*, l'objectif de PMI-IR est de choisir un synonyme parmi une liste donnée. Ces choix, notés $choix_i$, correspondent aux questions du TOEFL. Ainsi, le but est de calculer, pour chaque mot, le synonyme $choix_i$ qui donne le meilleur score. Pour ce faire, l'algorithme PMI-IR utilise différentes mesures fondées sur la proportion de documents dans lesquels les deux termes sont présents. Nous donnons ci-dessous une des mesures de base fondée sur l'Information Mutuelle utilisée dans les travaux de (Turney, 2001).

$$score(\ choix_i\ ) = \frac{nb(\ mot\ NEAR\ choix_i\ )}{nb(\ choix_i\ )}$$

- *nb(x) calcule le nombre de documents contenant le mot x,*
- *NEAR (utilisé dans la rubrique «recherche avancée» d'Altavista) est un opérateur qui précise si deux mots sont présents ensemble dans une fenêtre de 10 mots.*

Ainsi, la formule ci-dessus calcule la proportion de documents contenant *mot* et $choix_i$ dans une fenêtre de 10 mots par rapport au nombre de documents contenant le mot $choix_i$. Plus la proportion de documents contenant ces deux mots dans une même fenêtre est importante et plus *mot* et $choix_i$ sont considérés comme synonymes. D'autres formules plus élaborées ont également été appliquées. Ces formules utilisent les informations sur la présence de négations dans les fenêtres de 10 mots. Par exemple, les mots « grand » et « petit » ne sont pas synonymes si, dans une fenêtre, la présence d'une négation associée à un des deux mots est relevée.

Les mesures Web donnent une indication de popularité des associations de mots tout à fait intéressante lorsque des données issues d'un domaine plus ou moins général sont traitées. Par ailleurs, l'avantage de ces connaissances « externes » au corpus (c.-à-d.Web) tient au fait que nous sommes moins sensibles à la taille des données traitées (c.-à-d. corpus). En effet, cette taille et donc la fréquence d'apparition doit être assez significative lorsque des méthodes statistiques sont appliquées. Avec nos approches de type « Fouille du Web », nous n'avons pas de telles contraintes liées à la fréquence d'apparition des éléments dans les corpus eux-mêmes.

Dans cet article, nous proposons des mesures de Fouille du Web fondées sur des mesures statistiques afin d'identifier des EN (section 3) et leur catégorie (section 4). Ces étapes peuvent être exécutées séquentiellement dans le cadre d'un processus global de gestion des EN. Par exemple, dans un premier temps, nos approches permettent de déterminer que *hewlett packard* est une EN (entreprise fondée par Bill Hewlett and Dave Packard). La seconde étape peut alors mettre en exergue que cette EN est de type Organisation.

3. Fouille du Web pour l'identification des EN

Les EN sont peu sujettes aux variations (Fort *et al.*, 2009) telles que les «variations prépositionnelles». Nous allons nous appuyer sur cette constatation pour identifier les EN nominales à partir d'une liste de termes de type Nom-Nom dans un processus d'extraction de la terminologie (Roche et Kodratoff, 2009). Une telle liste est obtenue après l'application d'un processus de fouille de textes consistant, dans un premier temps, à normaliser puis étiqueter grammaticalement les textes. Ceci permet d'obtenir des listes de termes respectant des patrons syntaxiques définis (Nom-Nom, Nom-Préposition-Nom, Adjectif-Nom, etc.). Ensuite ces termes sont classés selon différentes méthodes statistiques.

Dans nos travaux sur les Entités Nommées, nous nous sommes focalisés sur l'exploitation des termes de type Nom-Nom. Le but consiste alors à identifier parmi les termes fournis par notre système, ceux représentant des Entités Nommées (Roche, 2011). Notons que notre approche n'utilise pas d'informations lexicales (présence de lettres en majuscule par exemple). En effet, notre méthode se veut adaptée au traitement de textes non normalisés (mails, blogs, textes ou fragments de textes intégralement en majuscule ou minuscule, etc). L'ajout de ces information permet bien sûr d'améliorer l'identification des EN.

Le processus d'identification des EN qui se décline en 3 étapes est résumé ci-dessous :

Etape 1 - Construction

Nous allons dans un premier temps construire des termes prépositionnels en nous appuyant sur quelques règles propres aux termes variants. De manière concrète, une forme variante d'un terme de type Nom-Nom en anglais (par exemple, *knowledge discovery*) est typiquement constituée d'une préposition associée à une permutation entre les noms (par exemple, *discovery of knowledge*).

Pour illustrer ce principe, considérons le terme *hewlett packard*. L'étape de construction permet d'obtenir le syntagme prépositionnel *packard of hewlett*. Celui-ci n'est clairement pas pertinent.

L'étape suivante du processus a alors pour but d'identifier les termes prépositionnels non pertinents afin de les considérer comme des EN.

Etape 2 - Fonction de Rang

Le but de la deuxième étape est de mesurer la dépendance entre chaque mot composant les termes construits. Pour calculer la dépendance, nous allons utiliser le principe donné en section 2 qui diffère de l'approche de Peter Turney selon deux axes :

1. utilisation d'une mesure statistique différente (mesure de Dice),
2. utilisation d'opérateurs différents (recherche exacte).

Pour mesurer la dépendance entre les mots (x, y, z) du terme prépositionnel construit, nous allons nous appuyer sur le coefficient de Dice (Smadja, 1993). Le choix de cette mesure est motivé par son bon comportement que nous avons montré dans nos précédents travaux (Roche et Kodratoff, 2009). Une telle mesure est définie par la formule suivante :

$$Dice_{\cap}(x, y, z) = \frac{3 \times nb(x \cap y \cap z)}{nb(x) + nb(y) + nb(z)}$$

Le cœur du calcul du numérateur de cette mesure consiste à évaluer si les mots (x, y, z) composant le syntagme construit sont strictement voisins (on effectue une recherche exacte, par exemple en utilisant les guillemets des moteurs de recherche). Dans notre exemple, de manière concrète, le calcul du numérateur de la formule relatif à notre exemple précédant correspond à effectuer la requête ''packard of hewlett''.

Etape 3 - Sélection

Les termes de type Nom-Nom qui obtiennent de faibles scores représentent des éléments peu enclins à la variation. Dans notre approche, de tels termes seront considérés comme des EN. Dans ce cadre, nous allons introduire un paramètre *S* qui représente un seuil de sélection. Par exemple, avec un seuil *S*=10, les dix termes ayant les scores les plus faibles seront prédits comme EN potentielles.

Après avoir présenté une méthode de Fouille du Web pour l'identification d'EN, la section suivante utilise des principes proches afin de déterminer la catégorie des EN (Personnes, Lieux, Organisations).

4. Fouille du Web pour la catégorisation des EN

L'approche que nous décrivons ci-dessous s'appuie dans un premier temps sur deux mesures qui calculent une certaine forme de dépendance entre le terme à catégoriser et des mots donnés (dépendance sémantique en section 4.1 et dépendance syntaxique en section 4.2). Puis deux types de combinaisons entre ces mesures seront proposées en sections 4.3 et 4.4.

4.1. Mesure SeScore

Le but de notre tâche est de mesurer la dépendance sémantique entre *x* et *y*. *x* représente l'EN à catégoriser et *y* représente un mot défini qui est associé à une catégorie d'EN donnée. En considérant l'exemple décrit en section précédente, le principe développé ici consiste à calculer une certaine forme de dépendance entre l'EN *hewlett packard* et un mot du domaine de la classe Organisation (par exemple, *company*).

Pour calculer la dépendance, nous allons nous appuyer sur la mesure de Dice associée à l'opérateur *Near* :

$$Dice_{Near}(x, y) = \frac{2 \times nb(x \text{ NEAR } y)}{nb(x) + nb(y)}$$

Cette mesure est alors prise en compte dans le cadre d'une mesure globale *SeScore* (*Semantic Score*).

Pour chaque catégorie *Cat* d'EN (Personne, Lieu, Organisation) un score est calculé. La catégorie qui obtient la valeur la plus élevée sera considérée comme la catégorie d'EN correcte. Pour valuer chaque catégorie, la mesure de Dice est calculée entre le mot *M* à catégoriser et des mots donnés qui sont associés à une classe d'EN. La moyenne des valeurs des mesures de *Dice* de chaque classe est prise en compte dans la mesure globale *SeScore* définie ci-dessous :

$$SeScore^{Cat}(M) = \frac{\sum_{i=1..N} Dice_{Near}(Se_i^{Cat}, M)}{N}$$

Dans cette mesure *SeScore*, *M* désigne le mot à catégoriser et Se_i^{Cat} représente un mot donné associé sémantiquement à une classe d'entité nommée *Cat*. Par exemple, nous utilisons les ensembles de mots Se_i^{Cat} ci-dessous pour chaque catégorie *Cat*.

- $Se^{Personne} = \{born\}$
- $Se^{Lieu} = \{city, country\}$
- $Se^{Organisation} = \{company, organisation, group\}$

Ainsi, le principe de cette mesure est de considérer la catégorie d'EN qui maximise *SeScore.* Le score calculé correspond à la moyenne des scores *de* $Dice_{Near}$ pour chaque catégorie *Cat*. Par exemple, en considérant le mot *M=Slovenia* à catégoriser, les deux calculs ci-dessous sont effectués :

$$Dice_{Near}(\text{city}, \text{Slovenia}) = \frac{2 \times nb(\text{city NEAR Slovenia})}{nb(\text{city}) + nb(\text{Slovenia})}$$

$$Dice_{Near}(\text{country}, \text{Slovenia}) = \frac{2 \times nb(\text{country NEAR Slovenia})}{nb(\text{country}) + nb(\text{Slovenia})}$$

Ceci permet de calculer la moyenne de ces deux scores en obtenant la valeur[1] ci-dessous avec le moteur de recherche Exalead qui propose la fonction NEAR :

$Se^{Lieu}(Slovenia)$=10103.

1 Cette valeur a été multipliée par un coefficient (10000000) pour améliorer l'interprétation dans cet article.

Notons que les valeurs obtenues pour $Se^{Personne}(Slovenia)$ et $Se^{Organisation}(Slovenia)$ sont respectivement 1964 et 2076. Ceci confirme donc que la mesure *SeScore* prédit correctement que l'EN *Slovenia* appartient à la catégorie Lieu.

Après avoir décrit cette mesure qui s'intéresse au calcul de **dépendance sémantique**, la section suivante se focalise sur une mesure, dont le principe est assez proche, mais qui s'intéresse à la **dépendance syntaxique**.

4.2. Mesure SyScore

Dans la suite, nous allons nous appuyer sur les structures syntaxiques pour aider à identifier la classe d'une entité nommée. En effet, selon les constructions syntaxiques, une EN peut être clairement associée à une classe. Par exemple, l'EN « *Montpellier* » peut être associée à une localisation si nous la retrouvons souvent dans une structure de type : « *I am in Montpellier* ». Bien sûr quelques contre-exemples existent avec des EN qui se retrouvent dans ce type de structure syntaxique (par exemple, « *I am in Microsoft Word* », où l'entité « *Microsoft Word* » est associée à la classe Logiciel). Cependant, statistiquement, au regard du nombre de pages sur le Web, la dépendance entre le mot « *in* » et une EN est beaucoup plus importante lorsque l'EN représente un lieu. Le principe de notre approche qui s'appuie sur cette hypothèse est détaillé ci-dessous.

La mesure *SyScore* (*Syntactic Score*) calcule la dépendance entre un mot donné (EN à catégoriser) et un mot fonctionnel qui forment un syntagme. Cependant, contrairement à la dépendance sémantique décrite en section 4.1, *SyScore* utilise la mesure de $Dice_{\cap}$ qui s'appuie sur une recherche exacte à deux éléments :

$$Dice_{\cap}(x, y) = \frac{2 \times nb(x \cap y)}{nb(x) + nb(y)}$$

Par exemple, « *in* » peut représenter un mot fonctionnel et « *Montpellier* » peut représenter le candidat EN afin de constituer le syntagme « *in Montpellier* ». Notre calcul de dépendance identifie dans quelle proportion un mot et un mot fonctionnel sont strictement voisins sur le Web. Dans ce cadre, le numérateur de la mesure de Dice est fondé sur le nombre de pages vérifiant la requête $x \cap y$ qui correspond aux pages où le mot fonctionnel x précède y.

Dans la mesure ci-dessous (*SyScore*) M désigne le mot à catégoriser et Sy_i^{Cat} représente un mot donné associé syntaxiquement à la catégorie d'entité nommée *Cat*.

$$SyScore^{Cat}(M) = \frac{\sum_{i=1..N} Dice_{\cap}(Sy_i^{Cat} \cap M)}{N}$$

Par exemple, pour un tel calcul, nous utilisons les données suivantes :

- $Sy^{Personne} = \{with\}$
- $Sy^{Lieu} = \{in\}$
- $Sy^{Organisation} = \{the\}$

Ainsi, le principe de cette mesure est de considérer la catégorie d'EN *Cat* qui maximise *SyScore.* Ce score correspondant à la valeur moyenne de $Dice_{\cap}$ de chaque catégorie.

En reprenant l'exemple de la section précédente (*M* = *Slovenia*), nous obtenons les valeurs suivantes :

- $SyScore^{Personne}$ (*Slovenia*) = 208
- $SyScore^{Lieu}$ (*Slovenia*) = 9442
- $SyScore^{Organisation}$ (*Slovenia*) = 99

Ceci confirme donc que la mesure *SyScore* fondée sur les informations syntaxiques prédit correctement que l'EN *Slovenia* appartient à la catégorie Lieu.

Après avoir détaillé chaque mesure, nous décrivons ci-dessous les différentes combinaisons possibles entre *SeScore* et *SyScore*.

4.3. Mesure globale par combinaison paramétrée

La première combinaison appelée *SySe* que nous proposons est une combinaison linéaire entre les deux mesures *SeScore* (section 4.1) et *SyScore* (section 4.2).

$$SySe_{\lambda}^{Cat}(M) = \lambda.SyScore^{Cat}(M) + (1-\lambda).SeScore^{Cat}(M)$$

Le paramètre $\lambda \in [0,1]$ donne un poids plus ou moins important à chacune des mesures.

Notons qu'en posant, λ=0, la mesure *SySe* revient à calculer la mesure *SeScore* et avec λ=1, la mesure *SySe* est assimilable à la mesure *SyScore*.

Rappelons que l'intérêt de cette mesure est de déterminer la catégorie *Cat* donnant la valeur qui maximise le score *SySe* pour une valeur λ donnée.

4.4. Mesure globale par combinaison et votes

Le défaut de la mesure précédente est lié au fait qu'elle est dépendante du paramètre λ à fixer. Pour pallier ce défaut nous proposons un système de vote qui prend en considération la classe majoritaire obtenue.

Le principe de cette approche appelée *VoSySe*, est de considérer la classe d'EN comme celle prédite par la majorité des classes données par *SySe* selon les valeurs de $\lambda \in [a,b]$. Les valeurs *a* et *b* seront discutées dans la section 5 ce cet article.

$$VoSySe_{\lambda}(M) = \{Cat/\{Vote_{\lambda=a...b}\{\max_{Cat}\{SySe_{\lambda}^{Cat}(M)\}\}\}\}$$

Concrètement, dans la mesure ci-dessus, *max* désigne la classe donnant la valeur maximum de *SySe* pour chaque $\lambda \in [a,b]$. La fonction *Vote* permet de considérer la classe majoritairement prédite (Personne, Lieu, Organisation).

Notons que si le nombre de classes prédites est égal, nous prendrons en compte un élément de plus (en considérant λ=*a-1* ou λ=*b+1*).

La section suivante donne, entre autres, les résultats de nos expérimentations selon les différentes valeurs de λ propres à la mesure *SySe* et les fenêtres de vote pour le système *VoSySe*.

5. Expérimentations

5.1. Evaluation de l'identification des EN

Afin d'évaluer la qualité de notre approche d'identification d'Entité Nommées, nous nous appuyons sur un corpus anglais composé de termes de type Nom-Nom. Cette liste contient 105 EN (issus des domaines de politique, militaire, religieux, etc) et 200 termes spécifiques (du domaine de la fouille de données).

Nous avons alors classé les 305 termes en appliquant notre fonction de rang $Dice_{\cap}$. Ces expérimentations ont nécessité l'exécution de 915 requêtes (3 requêtes par mesure).

Les résultats détaillés dans (Roche, 2011) montrent que 83% des premiers termes retournés par notre système (en posant S=40) sont correctement prédits comme des EN. Notons qu'une prédiction aléatoire donne un résultat autour de 34%. Ceci confirme la pertinence de notre approche sur des corpus en anglais, également testée sur des corpus en français. Notre système pourrait être amélioré en utilisant d'autres types d'informations : informations lexicales, informations sur le contexte d'où proviennent les termes, etc.

La section suivante évalue les méthodes proposées pour identifier les catégories d'EN.

5.2. Evaluation de la catégorisation des EN

5.2.1. Protocole expérimental

Dans ces expérimentations, 186 entités nommées ont été expertisées (62 EN dans chaque catégorie). Elles sont issues en grande majorité de la conférence CoNLL'2003 (Tjong Kim Sang et De Meulder, 2003). Elles ont été choisies aléatoirement à l'aide d'un programme prévu à cet effet.

La moitié des EN sont issues de données anglophones, l'autre moitié provient d'EN de données allemandes et hollandaises afin de mesurer la générécité de notre approche.

Ces expérimentations ont nécessité la mise en place de 1869 requêtes avec le moteur de recherche Exalead (http://www.exalead.com/) qui propose l'opérateur NEAR (mots placés dans une fenêtre de 16 mots). En effet, pour chaque EN à catégoriser, les mesures $Dice_{Near}$ et $Dice_{\cap}$ nécessitent l'exécution de 10 requêtes. Nous devons effectuer 9 requêtes pour calculer les numérateurs des mesures de *Dice* (requêtes associant les 9 mots issus de Se^{Cat} *et* Sy^{Cat} et le mot à catégoriser), une seule requête utile pour tous les dénominateurs (requête qui calcule le nombre de pages retournées avec le mot à catégoriser). Ceci demande donc l'exécution d'un total de 1860 requêtes. Les 9 requêtes supplémentaires sont exécutées une seule fois pour l'ensemble des expérimentations et correspondent au nombre de pages des mots issus de Se^{Cat} *et* Sy^{Cat}. Ces données seront utilisées pour l'ensemble des dénominateurs des mesures de *Dice*.

Après avoir détaillé le protocole expérimental, nous donnons ci-dessous les résultats obtenus dans le cadre de la mesure *SySe* qui combine les mesures *SyScore* et *SeScore*.

5.2.2. Evaluation de SySe

Le tableau ci-dessous montre la qualité de classification de manière globale (dernière colonne de notre tableau) et par catégorie. Globalement, la meilleure prédiction obtient une valeur de près de 69% avec λ=0.6 et λ=0.7. Rappelons qu'une classification aléatoire fournit une prédiction de 33%.

La prédiction est cependant inégale selon les catégories. En effet, les prédiction des lieux et des personnes ont un excellent comportement (de 80.6% à 83.9% dans les meilleures configurations). Cependant, la détection de l'organisation est plus complexe. Ceci peut être dû au fait que le nom des organisations (entreprises, mouvements politiques, organisations publiques, institutions politiques, etc.) est souvent ambigu et peut également être associé à un nom de lieu ou de personne.

Le tableau ci-dessous met également en exergue que les valeurs extrêmes (λ=0 et λ=1) donnent des prédictions globalement très faibles. Ceci signifie que la seule information sémantique ou syntaxique est souvent insuffisante ; les combinaisons se révèlent beaucoup plus fiables.

λ	EN Personnes (%)	EN Lieux (%)	EN Organisation (%)	**EN Total** (%)
0	**80.6**	67.7	29.0	59.1
0.1	**80.6**	71.0	30.6	60.7
0.2	**80.6**	74.2	38.7	64.5
0.3	**80.6**	75.8	40.3	65.6
0.4	**80.6**	80.6	40.3	67.2
0.5	**80.6**	82.3	41.9	68.3
0.6	**80.6**	82.3	43.5	**68.8**
0.7	77.4	**83.9**	45.2	**68.8**
0.8	72.6	82.3	45.2	66.7
0.9	69.3	80.6	**46.8**	65.6
1	22.6	79.0	45.2	48.9

5.2.3. Evaluation de VoSySe

Après avoir évalué la qualité de la combinaison (approche ***SySe***), nous allons expérimenter le système de votes ***VoSySe*** proposé en section 4.4.

Le tableau ci-dessous montre la qualité de la prédiction selon différentes fenêtres de λ. Ce tableau indique également le rang de la prédiction obtenu comparativement à ***SySe*** (cf. section précédente). Par exemple, la valeur « *(5)* » indique que le pourcentage correct de prédiction retourné par le système ***SySe*** est la 5ème meilleure valeur.

Ce tableau montre que le paramètre (fenêtre pour λ) le plus adapté consiste à ne considérer que la moitié supérieure (λ≥n/2).

Ces résultats montrent également que la prise en compte des extrémités (λ=0 et λ=1) n'influence pas les résultats globaux. En effet, l'intérêt de notre approche est lié au fait qu'un résultat localement peu satisfaisant a globalement moins d'influence dans un système de vote.

$VoSyS_\lambda$	EN Personnes (%)	EN Lieux (%)	EN Organisation (%)	**EN Total** (%)
$\lambda \in [0,n]$	**80.6** (1)	82.3 (2)	41.9 (4)	68.3 (2)
$\lambda \in [1,n-1]$	**80.6** (1)	82.3 (2)	41.9 (4)	68.3 (2)
$\lambda \in [n/2,n]$	77.4 (2)	**83.9** (1)	**45.2** (2)	**68.8** (1)
$\lambda \in [0,n/2]$	**80.6** (1)	75.8 (5)	40.3 (5)	65.6 (5)

5.3. Discussion : évaluation globale du processus

Il est à l'heure actuelle difficile d'évaluer le processus global en s'appuyant sur les critères de précision, rappel et F-mesure. En effet, les expérimentations issues de la première étape (identification des EN) ont consisté à extraire des types d'EN beaucoup plus larges que les trois catégories considérées dans la deuxième étape de notre processus (catégorisation des EN). Par exemple, à partir d'un corpus de CV en français, les EN nommées de type «logiciel» ont été extraites (par exemple, *lotus note*, *ciel paie*, *front page*, *ciel gestion*, etc). Ce type de catégorie n'est pas encore pris en compte dans la seconde étape du processus. De manière plus globale, dans nos futurs travaux, nous envisageons de considérer les classes d'EN plus riches qui sont proposées par le programme ACE (Doddington *et al,* 2004). Pour ce type d'EN, il sera nécessaire d'introduire de nouveaux critères sémantiques et syntaxiques.

Bien que le processus ne soit pas évalué de manière globale, chaque étape a été rigoureusement testée. Ainsi, les meilleurs résultats en terme de F-mesure pour l'identification des EN sont de 0.48 (précision = 0.35, rappel = 0.78) sur le corpus français de CV et de 0.66 (précision = 0.58, rappel = 0.77) sur un corpus anglais (Roche, 2011). Par ailleurs, les résultats de la catégorisation des EN à partir de benchmarks multilingues sont largement détaillés dans cet article. Prochainement, nous souhaitons rassembler ces différentes données afin de proposer une évaluation plus globale du processus.

6. Conclusions et perspectives

Les approches de type Fouille du Web peuvent se révéler particulièrement pertinentes pour identifier et catégoriser les Entités Nommées. Les approches proposées s'appuient sur diverses informations issues du Web.

Ces approches donnent des résultats parfaitement satisfaisants. Elles n'utilisent aucune connaissance endogène (propres aux candidats), ni de connaissances contextuelles.

Pour améliorer les résultats, nous proposons d'utiliser les informations liées aux conditions initiales dans lesquelles les candidats ont été extraits dans les corpus d'origine. Par exemple, il peut se révéler crucial d'utiliser des informations lexicales. Les candidats capitalisés pourraient alors être privilégiés par nos approches de prédiction. Par ailleurs les informations contextuelles de manière générale et les marqueurs syntaxiques présents dans un contexte local peuvent

également se révéler déterminants pour prédire qu'un mot ou un groupe de mots représente une Entité Nommée. Toutes ces informations sont par exemple exploitées dans la majorité des approches développées dans la conférence/challenge CoNLL. Leur association aux méthodes de Fouille du Web proposées dans cet article est donc tout à fait prometteuse.

Références

Bonnefoy L., Bellot P., Michel B. (2011). Une approche non supervisée pour le typage et la validation d'une réponse à une question en langage naturel : application à la tâche Entity de TREC 2010», *Proceedings of Conférence en Recherche d'Informations et Applications (CORIA)*

Brun C. and Hagège C. (2004). Intertwining deep syntactic processing and named entity detection. In *Proceedings of Advances in Natural Language Processing, 4th International Conference (EsTAL)*, p. 195-206

Cacheda, F., V. Carneiro, D.F., and Formoso, V. (2010). Performance evaluation of large-scale information retrieval systems scaling down. *Proceedings of the International Workshop on Large-Scale and Distributed Systems for Information Retrieval - SIGIR*

Carreras X., Màrquez L., and Padro L. (2003). A Simple Named Entity Extractor using AdaBoost. *Proceedings of Conference on Natural Language Learning*

Daille B., Fourour N., and Morin E. (2000). Catégorisation des noms propres : une étude en corpus, *Cahiers de Grammaire*, Vol 25, p. 115-129

De Meulder F. and Daelemans W. (2003). Memory-Based Named Entity Recognition using Unannotated Data. *Proceedings of Conference on Natural Language Learning*

Doddington G., Mitchell A., Przybocki M., Ramshaw L., Strassel S., and Weischedel R. (2004). The Automatic Content Extraction (ACE) Program, Tasks, Data, and Evaluation. *Proceedings of LREC,* p. 837-840

Fort K., Ehrmann M., and Nazarenko A. (2009). Vers une méthodologie d'annotation des entités nommées en corpus. *Proceedings of TALN (Traitement Automatique des Langues Naturelles)*

Nadeau D. and Sekine S. (2007). A survey of named entity recognition and classification, *Linguisticae Investigationes*, 30(1), p. 3-26

Nouvel D. and Soulet A. (2011). Annotation d'entités nommées par extraction de règles de transduction. *Proceedings of Extraction et Gestion des Connaissances (EGC)*, p. 119-130

Paik W., Liddy E.D., Yu E., McKenna M. (1994). Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval, *Corpus Processing for Lexical Acquisition*, MIT Press, chap. 4

Ritter A., Clark S., Mausam, Etzioni O. (2011). Named Entity Recognition in Tweets: An Experimental Study. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. p. 1524-1534

Roche M. (2011). How Statistical Information from the Web can Help Identify Named Entities, *Proceedings of International Conference on Web Information Systems (WEBIST),* Session Web and Text Mining

Roche M. and Kodratoff Y. (2009). Text and web mining approaches in order to build specialized ontologies. *Journal of Digital Information*, 10(4), 2009

Smadja F. (1993). Retrieving collocations from text : Xtract. *Computational Linguistics*, 19(1) p. 143-177

Tjong Kim Sang E.F. and De Meulder F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of Conference on Natural Language Learning*

Turney P. (2001). Mining the Web for synonyms : PMI-IR versus LSA on TOEFL. Proceedings of the 12th European Conference on Machine Learning (ECML), LNCS, p. 491-502