L'utilisation des entités nommées pour l'expansion sémantique des requêtes Web

Bissan Audeh, Philippe Beaune, Michel Beigbeder

École Nationale Supérieure des Mines de Saint-Étienne-ENSM.SE/FAYOL
158 Cours Fauriel
42023 Saint-Etienne cedex 2,
{audeh, beaune, mbeig}@emse.fr

Résumé. Les entités nommées sont des éléments intéressants pour les applications fondées sur le Traitement du Langage Naturel. Dans le cas de la recherche d'information, les entités nommées sont largement employées par les utilisateurs du web dans les requêtes de recherche, soit pour définir un concept de base, soit pour décrire un autre concept dans la requête. Du côté du modèle de recherche, les entités nommées sont des éléments riches en information qui aident à mieux cibler les documents pertinents. Dans cet article, nous étudions l'avantage d'étendre les entités nommées dans la requête. L'idée est d'utiliser une technique d'expansion sémantique sur une ontologie générale (Yago) pour désambiguïser les entités nommées et pour trouver leurs différentes appellations que l'on intègre dans la requête en utilisant 3 approches : sac de mots, dépendance séquentielle, et concept clé. Nous mesurons l'efficacité de ces expériences en termes de précision et rappel, et nous étudions l'effet du rôle des entités nommées sur l'expansion. Nous concluons que l'expansion des entités nommées est une méthode simple qui améliore significativement la qualité de la recherche quand elle est comparée à un modèle de référence sans expansion. De plus, cette méthode est assez compétitive par rapport à l'approche pseudo retour de pertinence souvent utilisée pour l'expansion de la requête.

1 Introduction

L'expansion de la requête est une approche souvent utilisée en recherche d'information pour aider le modèle de recherche à mieux identifier les documents pertinents. Le succès de cette technique dépend du bon choix des termes d'expansion et la façon d'ajouter ces nouveaux termes à la requête. D'un côté, si les nouveaux mots clés apportés par l'approche d'expansion ne sont pas pertinents pour le besoin d'information, la possibilité de trouver des documents pertinents baisse, ce qui a un effet négatif sur le rappel et la précision. Pour cette raison, le contrôle de la qualité des termes d'expansion est une étape indispensable de l'expansion de la requête. D'un autre côté, un bon choix des termes d'expansion ne suffit pas pour garantir le succès de l'expansion si ces termes n'ont pas été intégrés correctement dans la requête.
Dans cet article, nous nous concentrons sur les entités nommées, c'est à dire les termes qui

représentent des personnes, des lieux ou des événements. Ces termes, riches en information, ont motivé plusieurs études dans des domaines liés au Traitement du Langage Naturel. Ces études s'intéressaient à reconnaitre les entités nommées dans un texte (Guo et al., 2009), à les désambiguïser (Hoffart et al., 2011), ou à les classifier (Nadeau et Sekine, 2007). Dans le domaine de la recherche d'information, les entités nommées sont souvent utilisées pour l'annotation (Kiryakov et al., 2004), l'indexation (Buizza, 2011) ou la recherche (Petkova et Croft, 2007). Des travaux récents (Guo et al., 2009) ont remarqué l'importance des entités nommées pour les requêtes des utilisateurs, car ils ont constaté que plus de 70% des requêtes web contiennent au moins une entité nommée. En revanche, ces travaux s'intéressaient plutôt à la classification des entités nommées de la requête sans considérer l'effet de ces éléments sur la performance du modèle de recherche. Dans notre article, nous étudions l'expansion des entités nommées dans la requête, notamment par différentes méthodes de reformulation. Notre idée est construite sur trois éléments : premièrement, l'utilisation des entités nommées dans les requêtes est fréquente et utile (Guo et al., 2009; Kiryakov et al., 2004). En deuxième lieu, certaines études ont constaté que les requêtes web contiennent souvent une seule entité nommée (Guo et al., 2009). La troisième élément est l'avantage potentiel de l'expansion de la requête qui peut améliorer significativement la performance d'un modèle de recherche. Nous commençons en section 2 par une présentation des études précédentes sur l'expansion sémantique des requêtes, avec un zoom sur le cas des entités nommées. En section 3 nous présentons notre approche de l'expansion sémantique des entités nommées dont nous détaillons les différents étapes. La section 4 est consacrée à l'explication et l'évaluation de nos expériences.

2 État de l'art

Le but de l'expansion de la requête est de trouver des nouveaux documents pertinents (*i.e.* améliorer le rappel), et de tirer les documents pertinents déjà trouvés par la requête initiale vers le haut de la liste des résultats (*i.e.* améliorer la précision). Nous pouvons diviser les approches d'expansion de requête en deux catégories : la première catégorie regroupe les approches qui dépendent de la collection de documents. Ces approches peuvent être locales, comme les méthodes de retour de pertinence, ou globales. La deuxième catégorie contient les méthodes fondées sur l'utilisation d'une ressource externe comme une ontologie. Traditionnellement, pour ces deux catégories, les entités nommées sont souvent traitées comme les autres termes de la requête. Les approches fondées sur la collection de documents utilisent des calculs statistiques sur les termes dans les documents sans un traitement spécifique concernant les entités nommées. Par ailleurs, les méthodes fondées sur une ressource externe sont confrontées à la difficulté de traitement des entités nommées surtout quand la ressource externe décrit mal ces éléments (Navigli et Velardi, 2003). Par exemple, l'utilisation de WordNet pour l'expansion de la requête n'a pas beaucoup de succès (Mandala et al., 1998) : l'une des raisons est que les entités nommées dans cette ressource sont souvent manquantes, pas à jour, ou n'ont pas (ou très peu) de synonymes dans leurs Synsets [1]. Ainsi, WordNet n'est pas suffisant pour désambiguïser ou étendre les requêtes qui contiennent des entités nommées. Malgré ces difficultés liées à l'expansion de la requête, d'autres études en recherche d'information confirment l'importances des entités nommés dans les requêtes. Par exemple, certaines recherches sur les

1. Un Synset dans WordNet est un concept regroupant plusieurs synonymes.

requêtes longues considèrent qu'une sous requête contenant une entité nommée est un bon candidat qui doit être considéré pour la reformulation (Huston et Croft, 2010; Kumaran et Carvalho, 2009). D'autres études, comme (Maxwell et Croft, 2013), proposent un algorithme pour classer des groupes nominaux identifiés dans la requête afin de les utiliser pour construire une nouvelle requête.
Récemment, la disponibilité de ressources riches et ouvertes comme Wikipedia, a permis certains travaux explicitement dédiés à l'étude de l'expansion des entités nommées. Par exemple, Xu et al. (2008) ont utilisé Wikipedia pour extraire des termes qui sont sémantiquement proches des termes de la requête, alors que Brandao et al. (2011) ont étudiés des approches basées sur les pages et les *Infobox* de Wikipedia pour retrouver des expansions des entités nommées. Comme Xu et al. (2008) et Brandao et al. (2011), notre travail s'intéresse aussi à l'expansion des entités nommés, par contre, nous faisons le point sur les méthodes d'intégration des termes d'expansion dans la requête, ce qui n'était pas abordé dans les travaux précédents. De plus, nous étudions l'effet du rôle des entités nommées sur l'expansion.

3 L'expansion sémantique des entités nommées

Dans cette section nous présentons notre approche d'expansion de la requête fondée sur les entités nommées. Nous précisons dans les sous-sections suivantes les étapes de notre approche en considérant le cas des requêtes Web et l'ontologie générique YAGO (Suchanek et Weikum, 2007) que l'on utilise à la fois pour la désambiguïsation et pour le choix des termes d'expansion. Cette ontologie est une combinaison de WordNet et des catégories de Wikipedia. Elle a l'avantage d'être une ressource riche en entités nommées, liées entre elles par une grande variété de relations sémantiques (Hoffart et al., 2011).

3.1 La désambiguïsation

La désambïguisation est un problème complexe, largement abordé en Traitement du Langage Naturel (Navigli, 2009). Alors que ce problème est implicitepment résolu par les méthodes locales d'expansion de la requête, il reste un vrai défi pour les approches fondées sur l'utilisation d'une ressource externe. D'un autre côté, la désambiguïsation des entités nommées est un domaine relativement récent. À notre connaissance, la première méthode qui a traité ce sujet a été proposée en 2006 : elle utilisait Wikipedia en combinaison avec une approche d'apprentissage supervisé pour la désambïguisation des entités nommées (Bunescu et Pasca, 2006). Depuis, Wikipedia est devenu une ressource importante et à jour des entités nommées. Il faut noter également que dans le cas des requêtes web, la tâche de désambiguïsation est compliquée, car ces requêtes sont souvent très courtes, et contiennent donc peu d'éléments de contexte. Dans cet article, nous utilisons l'approche de désambiguïsation proposée par (Hoffart et al., 2011). Cette approche utilise l'outil de Stanford NER (Named Entity Recognition) pour identifier les entités nommées dans une requête, puis elle applique une combinaison de trois techniques de désambiguïsation : le sens le plus utilisé (*popularity prior*), la similarité (syntaxe et surface commune) entre une entité nommée et le concept correspondant dans l'ontologie, et enfin le *graphe de cohérence* entre les concepts. Quand il n'y a pas de contexte, l'algorithme va choisir la référence la plus courante pour l'entité à désambiguïser. Ce comportement est cohérent avec un scénario de recherche sur le Web, où on peut considérer que si l'utilisateur ne met

Topic	Entité Nommée	Variations sémantiques
515	Alexander Graham Bell	"Alexander Gram Bell", "Aleck Bell", "The father of the deaf"
517	Titanic	"Jinx Titanic"
478	Baltimore	"Baltimore City", "City of Baltimore", Baltamore, Bmore, "Baltimore riots", Baltimoreans, "Charm city" Mobtown, "Charm City"

TAB. 1 – *Exemples de variations sémantiques trouvés dans YAGO pour des entités nommées.*

que l'entité nommée dans sa requête, il souhaite alors orienter sa recherche vers la référence la plus courante de cette entité.

3.2 Le choix des termes

Une fois choisie la référence unique de l'entité nommée dans Yago, la sélection des termes d'expansion peut être effectuée. Ce choix dépend de la relation sémantique qu'on souhaite utiliser. Dans l'ontologie Yago, les relations sémantiques qui lient un concept à son entourage sont nombreuses [2]. Ces relations dépendent de la nature de chaque concept : par exemple une ville peut être reliée à sa surface ou à son nombre d'habitants, alors que d'autres types de relations sont utilisés dans le cas d'une personne (sa date de naissance par exemple). Dans le cas de l'expansion de la requête, le choix de la bonne relation sémantique n'est pas une tâche facile, elle sera le sujet de prochaines études. Pour cet article nous considérons la relation "Label" qui existe pour tous les concepts. L'avantage de cette relation est qu'elle lie une entité nommée à ses différentes appellations que l'on va appeler des *synonymes* pour simplifier. Ces synonymes peuvent être des alternatives orthographiques du terme (Baltimore-Baltamore), ou complètement différents au niveau de la syntaxe mais sémantiquement identiques au terme original (Baltimore-Mobtown). Le tableau 1 présente trois exemples d'entités nommées de type différent (personne, objet et lieu) avec leurs synonymes. Dans ce tableau, on constate également qu'un synonyme peut être une autre entité nommée ("Aleck Bell") ou un groupe nominal ("The father of the deaf").

3.3 La reformulation de la requête

La plupart des modèles de recherche considère la requête utilisateur comme un sac de mots, que l'on modifie avec l'expansion de la requête en ajoutant de nouveaux termes avec ou sans pondération. Selon le modèle de recherche d'information utilisé, des alternatives au sac de mots peuvent être considérées pour représenter la requête. Dans notre travail nous utilisons le modèle de *query likelyhood* (Strohman et al., 2004) qui est le modèle par défaut de l'outil de recherche d'information Indri [3]. Ce modèle est souvent utilisé comme modèle de référence dans les études en recherche d'information, il est fondé sur une combinaison d'un réseau bayesien et d'un modèle de langue, il propose un langage de requête assez flexible. Son fondement

2. Yago contient 72 types de relations sémantiques (`http://www.mpi-inf.mpg.de/yago-naga/yago/statistics.html`).
3. `http://www.lemurproject.org/`

est de calculer les croyances pour chaque nœud dans le réseau, puis de combiner ces croyances selon l'opérateur utilisé. Les équations 1 et 2 présentent comment Indri combine les croyances (b_i) pour les opérateurs $\#combine$ et $\#weight$ que nous utilisons dans notre approche.

$$b_{\#combine} = \prod_{i=1}^{n} b_i^{\frac{1}{n}} \tag{1}$$

$$b_{\#weight} = \prod_{i=1}^{n} b_i^{(\frac{w_i}{\sum_{j=1}^{n} w_j})} \tag{2}$$

Le troisième opérateur que nous utilisons $\#syn$ est un opérateur virtuel, dans le sens où il n'est pas associé à une équation de combinaison. Son rôle est de signaler au modèle de recherche qu'il ne faut pas distinguer les occurrences de chacun des termes donnés dans la liste de synonymes. Le langage de requête d'Indri permet également, avec les opérateurs $\#N$ et $\#uwN$ *(unordered window)*, d'exprimer le nombre maximum de mots autorisés entre des termes dans les documents du corpus pour qu'une occurrence soit comptabilisé [4].
Nous expliquerons par la suite les trois approches que l'on utilise pour l'intégration des nouveaux termes dans la requête avec ces opérateurs. Pour simplifier la compréhension de ces approches, nous utilisons la requête démonstrative suivante (issue de TREC n° 455 après la suppression des mots vides) :

```
Jackie Robinson appear first game
```

que l'on considère comme la requête de base sans expansion.

3.3.1 Sac De Mots (SDM)

Dans le cadre de l'expansion de la requête, l'utilisation du principe "sac de mots" signifie l'ajout de nouveaux termes à la requête originale, ce qui peut être codé par l'opérateur $\#combine$. Ainsi, l'expansion de notre exemple démonstratif pourrait donner la requête suivante :

```
#combine(Jackie Robinson appear first game
#1(Jackie Robinson) #1(Jack Roosevelt Robinson))
```

Néanmoins, dans cette requête étendue tous les éléments constitutifs du ***#combine*** seront interprétés indépendamment les uns par rapport aux autres, et donc avec une même probabilité. Dans notre cas, il faudrait que #1(Jackie Robinson) et #1(Jack Roosevelt Robinson) aient, l'un *ou* l'autre, la même probabilité que chacun des autres termes de la requête étendue. L'opérateur $\#syn$ nous permet de le faire avec cette formulation :

```
#combine(Jackie Robinson appear first game
#syn( #1(Jackie Robinson) #1(Jack Roosevelt Robinson)))
```

3.3.2 La Dépendance Séquentielle (DS)

La dépendance séquentielle a été proposée par Metzler et Croft (2005). L'idée d'origine a été fondée sur la reformulation des requêtes longues en trois parties pondérées : la première

4. Pour plus de détails sur ces opérateurs, voir `http://www.lemurproject.org/lemur/IndriQueryLanguage.php`

partie contient la requête originale sous la forme d'un sac de mots, cette partie a le poids le plus important. La deuxième partie combine des fenêtres de taille 1 pour chaque paire de termes successifs. La troisième partie contient les mêmes paires de termes mais avec une taille de fenêtre égale à 4. Dans (Maxwell et Croft, 2013), les auteurs ont adapté cette approche en remplaçant les paires de termes par des groupes nominaux qu'ils choisissent à partir de la requête initiale [5] en utilisant leur algorithme PhraseRank.
Notre expansion de requête s'inspire de ces travaux. Nous reformulons les requêtes initiales en trois parties, en utilisant les mêmes poids que dans (Metzler et Croft, 2005) et (Maxwell et Croft, 2013). Nous suivons aussi l'approche de (Maxwell et Croft, 2013) pour calculer la taille des fenêtres des expressions dans la dernière partie :

- requête initiale (coefficient de pondération = 0.85)
- extension avec les synonymes (issus de YAGO) dans des fenêtres de taille 1 (coefficient de pondération = 0.1)
- extension avec les même synonymes dans des fenêtres d'une taille équivalant à quatre fois le nombre de mots du groupe nominal (coefficient de pondération = 0.05)

Ce qui donne la requête suivante pour notre exemple :

```
#weight(
0.85 #combine(Jackie Robinson appear first game)
0.10 #combine(#syn(#1(Jackie Robinson)
  #1(Jack Roosevelt Robinson) appear first game)
0.05 #combine(#syn(#uw8(Jackie Robinson)
  #uw12(Jack Roosevelt Robinson) appear first game))))
```

3.3.3 Le Concept Clé (CC)

La troisième approche, fondée sur le principe de concept clé de (Bendersky et Croft, 2008), contient deux parties. La première partie est la requête originale sous forme d'un sac de mots, elle a le poids le plus important comme c'est le cas dans l'approche de dépendance séquentielle. La deuxième partie contient des groupes nominaux choisis dans la requête. Dans (Bendersky et Croft, 2008), ces groupes nominaux sont pondérés à l'intérieur de cette deuxième partie en fonction du modèle probabiliste qui les a classés, mais comme dans (Maxwell et Croft, 2013), nous n'utilisons pas cette pondération. Nous obtenons alors :

```
#weight(
0.8 #combine(Jackie Robinson appear first game)
0.2 #combine(#syn(#1(Jackie Robinson) #1(Jack
Roosevelt Robinson) appear first game)
```

Les poids $(0.8, 0.2)$ sont ceux de (Bendersky et Croft, 2008; Maxwell et Croft, 2013).

3.4 L'algorithme final

L'Algorithme 1 illustre les principales étapes de notre expansion sémantique des entités nommées.

Tout d'abord, nous commençons par initialiser la requête originale de l'utilisateur en enlevant les mots vides pour limiter le bruit (ligne 2). Ensuite, le processus de désambiguisation

5. Ce travail s'intéresse aux requêtes longues uniquement

Input :
q : Requête originale
m : Choix de Reformulation
Output :
Q : Requête étendue

```
1  begin
2      Q = FiltreMotsVides(q);
3      N = desambiguisationDesEN(Q);
4      foreach n ∈ N do
5          candidats_n = LabelsAnglais(n.geId(), Yago);
6          candidats_n = SynonymesAccepts(candidats_n);
7          exp_n = UniciteFiltre(candidats_n, Q);
8          Q = reformuler(Q, n, exp_n, m);
9      end
10     return Q
11 end
```

Algorithme 1 : L'algorithme d'expansion des entités nommées.

de Hoffart et al. (2011) identifie les entités nommées et choisit un unique sens pour chacune d'entre elles (ligne 3). En ligne 5, nous utilisons YAGO pour obtenir des termes d'expansions pour chaque entité nommée. Ces candidats à l'expansion passent ensuite par deux filtres : le premier (ligne 6) enlève les termes bruités (ceux qui contiennent moins de 3 caractères), et le deuxième assure qu'un terme d'expansion n'existe pas déjà dans la requête (ligne 7). Avec cette liste filtrée de termes d'expansion, la procédure de reformulation est finalement appelée (ligne 8) avec comme paramètre l'approche que l'on souhaite utiliser pour la reformulation (SDM, DS, ou CC).

4 Évaluation

Alors que la majorité des approches de reformulation ou d'expansion de requête s'intéresse plutôt à la précision en mesurant le MAP, R-Prec ou P@x (Maxwell et Croft, 2013; Carpineto et Romano, 2012; Bendersky et Croft, 2008; Petkova et Croft, 2007), dans notre travail nous considérons de plus le rappel pour obtenir une meilleure caractérisation de nos résultats. Pour cela, nous utilisons deux groupes de mesures : les mesures de précision (MAP, P@10, P@20, R-Prec), et les mesures de rappel (R@30, R@200, R@dernierRang). Les tests statistiques que nous utilisons sont "T-test" et "Randomization test" comme recommandé par (Smucker et al., 2007). Par notre évaluation nous souhaitons répondre aux questions suivantes : Quelle est l'effet des différentes méthodes d'intégration de nouveaux termes (SDM, DS, ou CC) ? Notre approche améliore-t-elle les résultats du modèle de référence sans expansion ? En comparant avec une approche traditionnelle d'expansion de la requête, où se situe notre approche ?
Nous faisons nos expériences sur deux collections web : WT10G [6] et ClueWeb09 [7] de TREC. Nous n'utilisons que les besoins d'informations dont le "Title" contient au moins une entité

6. Wt10g contient 1 692 096 documents génériques.
7. Pour plus d'information sur ClueWeb09 voir `http://lemurproject.org/clueweb09/`.

	Stemming	topics	Nombre de Titles contenant des entités nommées
WT10G	Porter	451-550	26
ClueWeb09	Krovetz	51-200	36

TAB. 2 – *Des informations sur les collections et requêtes de tests.*

nommée (tableau 2). Ce choix a été fait manuellement pour garantir que les requêtes de test contiennent au moins une entité nommée. Pour l'indexation et la recherche on utilise Indri5.5 sans modifier les paramètres par défaut.

4.1 Expériences

Dans le tableau 3, nous constatons que l'intégration des termes d'expansion dans la requête a un effet très important sur les résultats. Nous observons que le fait d'intégrer les termes d'expansion en mode sac de mots (sans pondération) dans la requête dégrade toutes les mesures de performance. Alors qu'avec les approches DS et CC, la précision et le rappel ont bien été améliorés. Ces améliorations sont même statistiquement significatives pour les mesures R-Prec, R@30 et R@200 pour la collection Wt10g. Pour la collection ClueWeb09, les approches DS et CC améliorent toutes les mesures par rapport à la base. Les tests de significativité de ces améliorations sont plus souvent positifs pour les métriques de précision.
En plus de comparer notre approche au modèle de référence sans expansion, il est intéressant de le comparer également à une autre approche d'expansion. Dans cette étude, nous comparons nos résultats à ceux obtenus en utilisant une méthode de retour de pertinence aveugle (*Pseudo Relevance Feedback* PRF) souvent utilisée comme une approche de base dans les études sur l'expansion de la requête. Pour cela, le tableau 3 propose également les résultats de l'approche d'expansion PRF par défaut de Indri sur la collection Wt10G, où nous avons fixé à 10 le nombre de documents de retour de pertinence et à 10 le nombre de termes d'expansion. En observant ces résultats, nous remarquons que PRF a dépassé notre approche en MAP, alors que pour le rappel, plus on avance dans les rangs des documents trouvés plus notre approche obtient un meilleur rappel par rapport à PRF. D'ailleurs, si on regarde le rappel au dernier rang, PRF diminue même le rappel du modèle de base sans expansion. Cela signifie que PRF a perdu des documents pertinents par rapport à la base sans expansion alors que notre approche en a trouvé des nouveaux.

Finalement, nous nous intéressons à comprendre si le fait que l'entité nommée est l'objet principal de la requête ou non a un effet sur les résultats après son expansion. Pour cela, nous divisons (manuellement) les 26 requêtes de l'expérience WT10G en deux groupes : le groupe A signifie que l'entité nommée soit le sujet de base dans la requête [8] (ex."Alexander Graham Bell"), dans le groupe B nous mettons les cas où l'entité nommée n'est pas le sujet de base, c'est-à-dire que si on enlève les autres termes de la requête cette dernière n'a plus le même sens (ex. Mexican food culture). Le tableau 4 contient le nombre de requêtes améliorées/dégradées par notre approche d'expansion (avec CC comme une méthode de reformulation) pour chaque mesure d'évaluation et selon le rôle d'entité nommée. Étant donné que le nombre de requêtes de test dans notre cas est relativement petit, il est difficile de tirer des conclusions significatives en divisant cette collection encore en deux groupes. Néanmoins, on peut remarquer que le rôle

8. Les requêtes qui ne contiennent pas d'autres termes que l'entité nommée sont forcément dans ce groupe.

(a) Évaluation de la précision

		MAP	P@10	P@20	R-Prec
WT10G	Baseline	22.27	34.44	29.44	24.56
	PRF	**25.62****	**37.78**	**31.48**	25.55
	SDM	19.36	30.74	26.67	20.02
	DS	23.97	34.44	31.11**	**26.46****
	CC	23.77	34.81	30.93	26.32**
ClueWeb09	Baseline	06.94	13.06	11.81	10.06
	SDM	07.19	13.89	12.08	09.61
	DS	07.96*	15.56*	12.50	10.97
	CC	**08.00***	**16.67****	**12.64**	**11.11***

(b) Évaluation du rappel

		R@30	R@200	R@dernierRang
WT10G	Baseline	22.60	55.17	76.90
	PRF	**25.21****	57.86	74.91
	SDM	20.42	51.26	72.31
	DS	24.25**	57.71**	77.38
	CC	24.41**	**57.97****	**77.44**
ClueWeb09	Baseline	04.34	21.86	50.37
	SDM	03.96	20.82	48.49
	DS	**04.83**	**23.57***	**53.11***
	CC	04.76	23.01	52.89

TAB. 3 – *Les résultats de l'expansion sémantique des entités nommées en précision (a) et en rappel (b). * signifie que les résultats sont statiquement significatifs (p<0.05) pour t-test ou randomization test, ** veut dire que les résultats sont significative pour les deux tests.*

	Rôle A		Rôle B	
Count	9		17	
	+	−	+	−
MAP	5	2	8	5
P@10	1	2	2	1
P@20	4	1	2	2
R-Prec	5	2	4	2
R@50	5	1	6	1
R@100	4	1	5	2
R@dernierRang	5	0	4	4

TAB. 4 – *Nombre de requêtes (WT10G) améliorées (+) ou dégradées (-) par l'expansion sémantique des entités nommées en utilisant l'option CC de la reformulation pour chaque rôle (A et B).*

de l'entité nommée a peu d'effet sur le MAP quand ces entités sont étendues. Alors que pour le rappel, on constate que pour le groupe A, toutes les requêtes ont obtenu un R@dernierRang supérieur ou égal à celui de l'approche sans expansion, mais pour le groupe B, une requête a une chance sur deux d'améliorer ce rappel une fois étendue.

4.2 Discussion

Les résultats précédents montrent le potentiel de l'expansion sémantique des entités nommées. Bien que l'ajout de nouveaux termes par l'approche SDM ait un effet négatif sur la performance, les approches DS et CC réussissent à améliorer significativement plusieurs mesures d'évaluations. En revanche, il faut prendre en compte que nous n'avons utilisé aucune pondération des termes, alors que souvent la plupart des méthodes d'expansion travaillent essentiellement sur ce sujet surtout avec le modèle SDM.
Notre observation concernant les meilleurs résultats de l'approche PRF par rapport à notre approche au niveau de la précision, signifie que l'utilisation des documents de retour de pertinence aide à modifier le classement des documents de façon à tirer les documents pertinents vers le haut de la liste. En revanche, il est intéressant de savoir que sans avoir accès à ces documents riches en information, et malgré le peu de texte dans les requêtes de test, l'expansion des entités nommées a réussi à trouver plus de documents pertinents que PRF, ce dernier ayant même perdu des documents pertinents trouvés par le modèle sans expansion.
Finalement, concernant les deux catégories d'entités nommées (A et B), nous constatons une amélioration plus stable au niveau du rappel quand l'entité nommée est le sujet principal de la requête. Cette observation semble logique, car l'ajout de synonymes pour ce genre d'entités est probablement utile pour trouver plus de documents pertinents, surtout que dans ce cas il y a peu d'ambiguïté si on suppose que l'utilisateur pense au sens le plus courant quand il construit une requête qui ne contient que l'entité nommée.

5 Conclusion

Dans cet article, nous avons proposé une nouvelle idée d'expansion de la requête fondée sur les entités nommées. Cette idée a été le résultat d'une étude sur l'expansion sémantique de la requête en utilisant une ontologie, et sur l'importance des entités nommées dans les requêtes. Nos expériences ont montré des résultats encourageants sur deux collections web de TREC. L'approche que nous avons proposée a montré qu'en considérant d'autres méthodes d'intégration de termes que le traditionnel sac de mots, on peut significativement améliorer la précision et le rappel d'un modèle de référence, même sans utiliser de pondération des termes. Par la suite, nous souhaitons examiner le choix des relations sémantiques pour les entités nommées en fonction de leur nature et de leur contexte dans la requête, au lieu de prendre le "Label" pour toutes les entités nommées. De plus, nous pensons étudier la pondération des termes pour le modèle SDM afin qu'il soit comparable avec les autres méthodes.

Références

Bendersky, M. et W. B. Croft (2008). Discovering key concepts in verbose queries. In *SIGIR*. ACM.

Brandao, W. C., A. S. Silva, E. S. Moura, et N. Ziviani (2011). Exploiting entity semantics for query expansion. In *WWW/Internet*. IADIS.

Buizza, P. (2011). Indexing concepts and/or named entities. *JLIS. IT*.

Bunescu, R. et M. Pasca (2006). Using encyclopedic knowledge for named entity disambiguation. In *EACL*.

Carpineto, C. et G. Romano (2012). A Survey of Automatic Query Expansion in Information Retrieval. *ACM Computing Surveys*.

Guo, J., G. Xu, X. Cheng, et H. Li (2009). Named entity recognition in query. In *SIGIR*. ACM.

Hoffart, J., M. A. Yosef, I. Bordino, H. Furstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, et G. Weikum (2011). Robust Disambiguation of Named Entities in Text. In *EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Huston, S. et W. B. Croft (2010). Evaluating verbose query processing techniques. In *SIGIR*. ACM.

Kiryakov, A., B. Popov, I. Terziev, D. Manov, et D. Ognyanoff (2004). Semantic annotation, indexing, and retrieval. *Web Semantics : Science, Services and Agents on the World Wide Web*.

Kumaran, G. et V. R. Carvalho (2009). Reducing long queries using query quality predictors. In *SIGIR*. ACM.

Mandala, R., T. Takenobu, et T. Hozumi (1998). The use of WordNet in Information Retrieval. In *ACL Workshop on the Usage of WordNet in Information Retrieval*. ACL.

Maxwell, K. T. et W. B. Croft (2013). Compact query term selection using topically related text. In *SIGIR*. ACM.

Metzler, D. et W. B. Croft (2005). A markov random field model for term dependencies. In *SIGIR*. ACM.

Nadeau, D. et S. Sekine (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*.

Navigli, R. (2009). Word sense disambiguation : A survey. *ACM Computing Surveys*.

Navigli, R. et P. Velardi (2003). An analysis of ontology-based query expansion strategies. In *14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining*.

Petkova, D. et W. B. Croft (2007). Proximity-based document representation for named entity retrieval. In *CIKM*. ACM.

Smucker, M., J. Allan, et B. Carterette (2007). A comparison of statistical significance tests for information retrieval evaluation. In *CIKM*.

Strohman, T., D. Metzler, H. Turtle, et W. Croft (2004). Indri : A language-model based search engine for complex queries. In *International Conference on Intelligence Analysis*.

Suchanek, F. M. et G. Weikum (2007). YAGO : A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In *Proceedings of the 16th international conference on World Wide Web*.

Xu, Y., F. Ding, et B. Wang (2008). Entity-based query reformulation using wikipedia. In *CIKM*. ACM.

Summary

Named entities are interesting elements for applications based on Natural Language Processing (NLP). In the case of information retrieval, named entities are widely used in users web queries either to define a main concept or to describe other concepts in the query. On the other hand, named entities are also important for the retrieval model, because they are considered as rich information containers that help the model to better recognize relevant documents. In this paper, we study the advantage of expanding named entities in users' queries. We use a semantic query expansion method on the generic ontology YAGO, which is used to disambiguate and to find synonyms for named entities. We focus on the way in which new terms are integrated in the query, and the effect of the named entity role on the expansion. Three query reformulation approaches are explored: Bag Of Words, Sequential Dependence and Key Concepts. We measure the performance of theses approaches regarding Recall and Precision. We conclude that expanding named entities with a suitable reformulation approach can significantly enhances the performance of a baseline with no expansion. In addition, it is a competitive method when compared to a traditional expansion approach like pseudo relevance feedback.