Vers un modèle de langue mixte concepts-mots pour la recherche d'information

Lynda Said L'hadj*—Mohand Boughanem—Karima Amrouche***

** Ecole nationale Supérieure d'Informatique ESI, BP 68M, 16270, Oued Smar, Algérie*

{l_said_lhadj, k_amrouche}@esi.dz

*** Laboratoire IRIT, Université Paul Sabatier*

118 Route de Narbonne 31 062 Toulouse Cedex 09, France

Bougha@irit.fr

RÉSUMÉ. La majorité des modèles de langue appliqués à la recherche d'information repose sur l'hypothèse d'indépendance des mots apparaissant dans les documents et les requêtes. Plus précisément, ces modèles sont estimés à partir des mots simples sans considérer les éventuelles relations sémantiques et conceptuelles. Pour pallier ce problème, deux grandes approches ont été explorées : la première intègre des dépendances d'ordre surfacique entre les mots (bi-grammes, bi-termes), et la seconde repose sur l'utilisation des ressources sémantiques pour capturer les dépendances entre les mots. Le modèle de langue que nous présentons dans cet article s'inscrit dans la seconde approche. Nous proposons de lever la contrainte d'indépendance des mots par une représentation des documents et requêtes intégrant les concepts qu'ils recèlent.

ABSTRACT. The majority of language models applied to information retrieval is based on word independence hypothesis. More precisely, those models are estimated without considering semantic or conceptual relations between those words. To palliate this problem, two principal approaches have been explored: the first one integrates syntactic dependencies between words (bi-grams, bi-terms) and the second approach is based on the use of semantic resources to integrate words dependencies. The language model proposed in this paper is in the second approach. We propose to relax the independence terms constraint by representing both documents and queries with concepts.

MOTS-CLÉS : Recherche d'information, modèle de langue, concepts.

KEYWORDS: Information retrieval, language model, concepts.

1. Introduction

Depuis leur introduction en Recherche d'Information (RI), les modèles de langue se sont distingués par leur efficacité et leur fondement mathématique solide, (Ponte et al.,1998), (Song et al.,1998). Contrairement aux autres modèles de RI (des détails sur ces modèles peuvent être trouvés dans (Baeza et al., 1999)), les modèles de langue ne modélisent pas directement la notion de pertinence d'un document (D) face à une requête (Q). La pertinence d'un document vis-à-vis d'une requête est vue comme la probabilité que la requête soit générée par le modèle de langue du document soit $P(Q|D)$ appelée aussi *score de pertinence*. L'estimation de $P(Q|D)$ s'effectue sous l'hypothèse d'*indépendance des mots* qui simplifie grandement le calcul mathématique mais entraîne la représentation des documents et des requêtes comme des suites de mots (graphies) dénuées de sens. Par conséquent, le problème longtemps connu en RI classique de la non prise en compte des phénomènes de polysémie et de synonymie rebondit dans les modèles de langue.

Pour pallier ce problème et bonifier davantage les résultats des modèles de langue, une nouvelle génération de modèles a été développée. Elle s'inscrit à l'intersection des modèles de langue et de la recherche sémantique d'information (Alvarez et al., 2004), (Bao et al., 2006), (Bai et al., 2005) (Baziz, 2005), (Cao et al., 2005), (Gao et al., 2004), (Srikanth et al., 2002, 2003). Dans cette intersection, deux grandes approches peuvent être distinguées : *l'approche statistique surfacique* qui prend en compte des dépendances d'ordre surfacique entre les mots pour capturer les contenus sémantiques et *l'approche sémantique* qui prend en compte le sens des mots dans le modèle de langue. Cette seconde approche repose sur l'utilisation des ressources sémantiques (ontologies, thésaurus) pour identifier les sens des mots.

Le modèle de langue présenté dans cet article s'inscrit dans la seconde approche. Nous proposons de capturer les dépendances entre les mots par l'identification des concepts auxquels ils renvoient. Il a été constaté que l'approche conceptuelle souffre du problème de silence dû à la non disponibilité de ressources conceptuelles complètes et générales (Baziz, 2005). De plus, elle mélange les concepts génériques ainsi que les concepts spécifiques alors que ceux-ci ont des effets différents sur la performance d'un modèle de RI. Nous proposons alors un modèle de langue mixte qui tient compte non seulement des concepts détectés dans l'ontologie (WordNet) mais aussi des concepts non détectés dans l'ontologie. L'utilisation de l'ontologie permet aussi d'intégrer les dépendances entre concepts et de séparer les concepts de niveau générique des concepts de niveau spécifique.

Le reste du papier est organisé comme suit : dans la section 2, nous présentons un tour d'horizon des travaux qui se situent à l'intersection des modèles de langue et de la recherche sémantique d'information. La section 3 est consacrée à la présentation détaillée du modèle proposé, puis nous déroulons un exemple pour illustrer notre modèle dans la section 4. Enfin, dans la section 5, nous terminons ce papier avec une synthèse du travail présenté.

2. Etat de l'art

Généralement, dans les modèles de langue, la pertinence d'un document face à une requête est en rapport avec la probabilité que la requête Q puisse être générée par le modèle de langue du document M_D (Ponte et al.,1998). Le score de pertinence du document D face à une requête Q vue comme une suite de mots $Q = t_1 t_2 ... t_n$ est déterminé par :

$$Score(Q, D) = P(Q|M_D)^1 = P(t_1 t_2 ... t_n | M_D) \quad [1]$$

Pour estimer ce score [1], il faudrait que les mots t_i soient *indépendants* ainsi:

$$score(Q, D) = \prod_{i=1}^{n} (t_i \mid M_D) \quad [2]$$

Cette hypothèse d'indépendance des mots pose deux problèmes majeurs :

– **Données éparses** : dans les modèles de langue, si un mot t_i de la requête est absent dans le document, le score de pertinence est alors nul même si les autres mots $t_{j, j \neq i}$ sont présents.

– **Représentation en sac de mots** : Cette indépendance n'est à vrai dire pas adéquate avec la réalité du langage naturel, où les mots sont liés les uns aux autres. En d'autres mots, l'hypothèse d'indépendance des mots ne permet pas la prise en compte de deux phénomènes très importants en RI, à savoir la *polysémie* et la *synonymie*.

Le problème de données éparses a été résolu en introduisant les techniques de lissage (exemple : lissage par interpolation) (Stanley Chen et al., 1998). Ces dernières consistent à attribuer une probabilité non nulle aux mots de la requête absents dans le document. Elles ont également été utilisées pour résoudre le problème de représentation des documents et des requêtes en sac de mots, car elles offrent la possibilité de combiner plusieurs sources de termes, (Cao et al., 2005), permettant ainsi un lissage dit *sémantique*. De nombreux travaux ont été réalisés dans ce sens. Ils peuvent être classés en deux approches selon (Cao et al., 2005) : *L'approche guidée par l'apprentissage des données (approche statistique ou surfacique)* et *l'approche guidée par les ressources sémantiques.*

Les approches surfaciques tentent d'intégrer les relations entre les mots en se basant sur des considérations statistiques, par exemple, la cooccurrence entre les mots. L'un des premiers travaux à avoir introduit les dépendances surfaciques entre les mots est celui de (Berger et al., 1999), ils ont proposé un modèle de translation $t(q_i \mid w)$ [2] (formule [3]) qui traduit les liens potentiels entre un mot de la requête q_i et ceux du document w .

$$P(Q|D) = \prod_{i=1}^{n} \sum_{w} t(q_i \mid w) P(w \mid D) \quad [3]$$

---

[1] Dans ce qui suit, on écrira $P(Q|D)$ pour représenter la probabilité $P(Q|M_D)$.

[2] Le facteur $t(q_i|w)$ est estimé en utilisant une variante des cooccurrences de q_i et de w.

(Gao et al., 2004) ont modélisé les liens entre les mots en utilisant une variable cachée *L* (Linkage), laquelle est intégrée dans l'estimation de $P(Q|D)$. Cette variable est représentée par un graphe acyclique non orienté où sont considérées les dépendances les plus distantes et les plus solides entre les mots de la requête. Ceci réduit non seulement le temps d'exécution mais aussi les erreurs d'estimation des scores de dépendance dues aux données éparses.

De leur coté, (Srikanth et al., 2002) ont étendu le modèle bi-grammes à un modèle bi-termes où la contrainte d'adjacence et de l'ordre des mots est ignorée. Dans (Srikanth et al., 2003), les auteurs ont proposé un modèle uni-gramme de concepts où la requête est considérée comme une séquence de concepts $\{c_1,c_2,\ldots,c_k\}$, chaque c_j est une séquence de mots $\{q_1^j, q_2^j, \ldots, q_{i_k}^j\}$. La probabilité de vraisemblance de la requête est donnée par :

$$P(Q|M_D) = \prod_j P(c_j|M_D) \qquad [4]$$

Les concepts c_j identifiés après un étiquetage syntaxique sont supposés être indépendants.

Les modèles cités ci-dessus ont montré des résultats meilleurs que le modèle de langue uni-gramme. Cependant, nous pensons que ce type d'analyse statistique engendre beaucoup de bruit et il est quasiment impossible de l'éliminer sans filtrage linguistique voire sémantique. En outre, l'approche statistique (surfacique) ne peut capturer la sémantique implicite (synonymes, hyperonymes…) des requêtes et des documents et ne peut résoudre le problème de polysémie. C'est pourquoi une autre direction utilisant les ressources sémantiques a été suivie.

Cette nouvelle direction se base sur des liens sémantiques extraits de ressources sémantiques comme les ontologies. (Cao et al., 2005) ont proposé un modèle intéressant pour incorporer les dépendances entre les mots de la requête et ceux du document. Ils combinent le modèle uni-gramme classique avec le modèle de dépendance des mots. Dans ce dernier, ils intègrent des dépendances d'ordre statistique (cooccurrence) et sémantique (relations entre mots simples extraites de WordNet) en exploitant les techniques de lissage.

Dans le même objectif, (Bao et al., 2006) ont proposé un modèle de langue basé sur les sens des mots. Ils combinent un modèle de langue uni-gramme de mots simples avec un modèle de leurs sens respectifs identifiés par un système de désambiguïsation basé sur WordNet.

Les résultats des expérimentations des modèles des deux approches sont tous aussi prometteurs les uns que les autres. Ils traitent le problème de dépendance des mots dans les modèles de langue. Cette dépendance peut être de deux types : la dépendance entre les mots dans la requête (ou dans le document) ou les dépendances entre les mots de la requête et les mots du document (Cao et al., 2005). La majorité des modèles développés jusque là s'est intéressée au premier type de dépendance bien qu'en RI les deux types sont importants.

Pour notre part, nous proposons d'intégrer ces deux types de dépendance en identifiant les concepts que recèlent les requêtes et les documents en nous appuyant sur les ontologies (WordNet). En effet, nous pensons que la simple désambigüisation[3] des mots simples comme dans (Bao et al., 2006) ou bien l'utilisation des relations (surfaciques ou sémantiques) entre mots simples uniquement (Cao et al., 2006) ne suffit pas pour capturer le contenu sémantique des documents et des requêtes étant donnée l'ambiguïté du langage naturel. Nous pensons de ce fait qu'un concept, correspondant par exemple à une entrée dans une ontologie, est plus précis qu'un mot isolé ou un sens isolé. Nous proposons alors un modèle de langue basé sur les concepts et nous supposons qu'un concept peut être représenté par un mot simple où par un groupe de mots.

3. Modèle proposé

Le modèle de langue que nous proposons combine les concepts identifiés à travers une ontologie et ceux qui ne le sont pas. La technique utilisée pour la construction des descripteurs du document et de la requête est celle décrite par (Baziz, 2005). Ensuite, la pertinence du document face à la requête est estimée par la probabilité que le modèle de la requête (descripteur) soit généré par celui du document.

3.1. Construction du descripteur

Tout document (respectivement requête) est projeté sur une ontologie par exemple WordNet, les termes (mots simples ou groupe de mots) ayant une entrée dans l'ontologie sont alors pris comme des éléments du document. Pour ce faire, nous appliquons la méthode de (Baziz, 2005) (Boughanem & al., 2007) pour la construction de l'arborescence du document et de la requête (cette méthode inclut détection des groupes de mots, désambigüisation et pondération). Les termes non identifiés dans l'ontologie sont gardés dans le descripteur car il peut arriver que des concepts importants n'existent pas dans l'ontologie (cas des *noms propres* ou de *néologismes*).

3.2. Modèle de langue à base de concepts

Une fois que tous les concepts sont identifiés, nous considérons la requête et le document comme des sacs de concepts. Nous appelons concepts aussi bien les termes identifiés dans l'ontologie que ceux non identifiés. Ainsi: $Q= c_1c_2...c_n$.

$$P(Q \mid D) = \prod_{c_i \in Q} P(c_i/D) \qquad [5]$$

[3] Dans ce papier, nous entendons par désambiguïsation, l'identification automatique du sens d'un terme polysémique.

$P(c_i \mid D)$ est la probabilité qu'un concept c_i de la requête soit généré par un concept du document. Ce concept peut alors être généré *directement* par le document, c'est-à-dire c_i est effectivement présent dans le document, ou bien *indirectement* par un concept non présent mais sémantiquement lié au concept c_i. Nous tenons compte de ce fait en proposant d'estimer cette probabilité en utilisant la technique de lissage par interpolation linéaire (Il s'agit d'un lissage sémantique à proprement parler).

$$P(c_i/D) = \lambda P_{\overline{exp}}(c_i/D) + (1-\lambda)P_{exp}(c_i/D) \quad [6]$$

$P_{\overline{exp}}(c_i/D)$: est la probabilité que c_i (ne correspondant à aucune entrée de WordNet ne peut être étendu d'où la notation $\overline{exp}$) soit généré par D.
$P_{exp}(c_i/D)$: est la probabilité que le concept c_i (correspondant à une entrée de WordNet) soit généré par D (i.e. c_i peut être étendu selon la relation de subsomption).

Estimation de $P_{exp}(c_i \mid D)$

Cette probabilité représente le modèle qui permet de représenter les concepts de la requête présents dans le document directement (appariement direct) ou indirectement (appariement indirect). Nous proposons alors de combiner ces deux types d'appariement dans un modèle d'interpolation linéaire :

$$P_{exp}(c_i/D) = \theta P_p(c_i/D) + (1-\theta)P_{\bar{p}}(c_i/D) \quad [7]$$

On note $P_p(c_i/D)$ la probabilité que le concept c_i soit généré directement par D, et $P_{\bar{p}}(c_i/D)$ est la probabilité que le concept c_i soit généré indirectement (par des concepts sémantiquement liés par le lien de subsomption) par le document D.

Estimation des probabilités $P_{\overline{exp}}(c_i/D)$ **et** $P_p(c_i/D)$

Les probabilités que c_i corresponde à un concept qui apparait explicitement dans la requête sont estimées, par analogie au modèle uni-gramme de mots simples, où :

$$P(t_i/D) = \frac{tf(t_i,D)}{\sum_{t_j \in D} tf(t_j,D)}$$

Nous adaptons cette probabilité pour les concepts, c'est-à-dire au lieu de considérer le mot simple t_i comme unité d'indexation, nous considérons le concept c_i. Nous appliquons la méthode de pondération des concepts de *CF (Concept Frequency)* (Baziz, 2005).

$$P_{\overline{exp}}(c_i/D) = P\left(c_{\overline{exp_i}}/D\right) = \frac{cf(c_{\overline{exp_i}},D)}{\sum_{c_j \in D} cf(c_j,D)} \quad [8]$$

Où $cf\left(c_j, D\right) = Count(c_j, D)$ qui retourne la fréquence d'apparition de $c_{\overline{exp_i}}$ (concept non identifié dans WordNet) dans le document.

$$P_p(c_i/D) = P\left(c_{pi}/D\right) = \frac{cf(c_{pi},D)}{\sum_{c_j \in D} cf(c_j,D)} \quad [9]$$

$$cf(c_j, D) = Count(c_j, D) + \sum_{SC \in subconcept(cj)} \frac{length(Sc)}{Length\ (cj)}\ Count\ (c_j)$$

Où *Length(C)* représente le nombre de mots dans le concept *C* et *sub_concept (C)* le nombre de tous les sous-concepts[4] (qui doivent correspondre à leur tour à des concepts de l'ontologie) dérivés de *C*: sous-concepts de *n-1* mots de *C*, sous-concepts de *n-2*,.. et tous les mots uniques de *C*.

Il peut arriver que c_i (concept de la requête) identifié ou non dans WordNet soit absent dans le document. Dans ce cas P(Q | D) est nulle, même si le document contient les autres concepts, nous rebondissons ainsi sur le problème des probabilités nulles des modèles uni-gramme de mots simples. Alors nous lissons les probabilités $P_p(c_i \mid D)$ et $P_{\overline{exp}}(c_i \mid D)$ en utilisant la méthode dite "Absolute Discount", (Stanley Chen et al., 1998), que nous adaptons aux concepts :

$$P_{abs}(c_i \mid D) = \frac{\max(cf(c_i,D)-\delta,0)}{|D|} + \frac{\delta |D|}{|D|} P_{MLE}(c_i \mid C)$$

Où $|D| = \sum_{c_j \in D} cf(c_j, D)$

$P_{MLE}(c_i \mid C)$ est le maximum de vraisemblance du concept c_i dans la collection C, elle est donnée par :

$$P_{MLE}(c_i \mid C) = \frac{cf(c_i, C)}{\sum_{c_i \in C} cf(c_i, C)}$$

Estimation de $P_{\bar{p}}(c_i \mid D)$

Cette probabilité représente le modèle d'expansion de la requête, où il s'agit de retrouver les documents contenant non seulement les mêmes concepts que la requête, mais aussi les concepts sémantiquement proches de ceux de la requête.

Un concept peut être lié à plusieurs concepts appartenant à un ensemble donné (sous hiérarchie de WordNet), alors, le processus d'expansion doit être limité, sinon on perd en précision, si on exploite toute la hiérarchie de l'ontologie. Alors nous proposons de limiter cette expansion sur le référentiel de représentation du document et requête proposé dans (Baziz, 2005). Pour tenir compte de ces concepts d'expansion, nous avons opté pour le modèle de translation statistique de (Berger et al., 1999) qui est le plus adéquat à intégrer les concepts liés à ceux de la requête.

$$P_{\bar{p}}(c_i \mid D) = \sum_{c_{\bar{p}} \in E} P(c_i \mid c_{\bar{p}})\, P(c_{\bar{p}} \mid D) \qquad [10]$$

E : est l'ensemble de tous les concepts liés à ceux de la requête.

Les approches conceptuelles proposées jusque là, mélangent concepts spécifiques et concepts génériques. En outre, les résultats des travaux de chercheurs en RI dont (Boughanem & al., 2007) et (Baziz, 2005) stipulent qu'il est important de préciser les sens de l'expansion en exploitant la relation "*is a*", car les concepts

[4] La formule (10) est une généralisation de la formule (9), quand un concept n'a pas de sous concepts ou n'existe pas dans l'ontologie le second terme est alors nul.

génériques améliorent le rappel tandis que les concepts spécifiques améliorent la précision, cependant, il faut limiter l'expansion à un certain niveau, sinon les performances du modèle se dégraderaient.

Les modèles de langue offrent la possibilité de séparer ces deux niveaux de concepts, alors nous considérons deux sources pour générer le concept de la requête c_i (en exploitant les sous arbres conçus lors de l'extraction des concepts). Il peut être généré par un concept plus spécifique (c'est-à-dire placé en dessous de c_i dans la hiérarchie de concepts de WordNet), ou par un concept plus générique (placé en dessus de c_i dans la hiérarchie de concepts de WordNet). Nous pouvons modéliser cette relation entre concepts par un lissage par interpolation linéaire. Ainsi :

$$P_{\overline{p}}(c_i \mid D) = \alpha \left[\sum_{\substack{c_g \in E \\ c_g \notin Q}} P(c_i \mid c_g)\, P(c_g \mid D) \right] + (1-\alpha) \left[\sum_{\substack{c_s \in E \\ c_s \notin Q}} P(c_i \mid c_s)\, P(c_s \mid D) \right] \quad [11]$$

Où : $P(c_i|c_g)$ est la probabilité conditionnelle que c_i soit généré par un concept générique c_g (hyperonyme). $P(c_i \mid c_s)$ est la probabilité conditionnelle que c_i soit généré par un concept plus spécifique c_s (hyponyme).
Nous appelons $\sum_{\substack{c_g \in E \\ c_g \notin Q}} P(c_i \mid c_g)\, P(c_g \mid D)$ modèle d'expansion générique pondéré avec le paramètre α, et $\sum_{\substack{c_s \in E \\ c_s \notin Q}} P(c_i \mid c_s)\, P(c_s \mid D)$ modèle d'expansion spécifique pondéré avec $(1-\alpha)$[5].

Estimation de $P(c_i \mid c_g)$ et de $P(c_i \mid c_s)$

Ces deux probabilités interprètent le degré de liaison (distance sémantique) entre c_i et c_g ou c_s. Pour les estimer, nous utilisons une mesure de similarité sémantique entre concepts basée sur la distance sémantique (basée sur le lien de subsomption).
Notre choix est fixé sur la mesure de (Wu et al., 1994), elle est simple à mettre en œuvre et a montré son efficacité selon (Lin, 1998). Elle est basée sur le principe suivant : Soient X et Y deux éléments d'une ontologie, la similarité entre eux est basée sur les distances N1 et N2 qui les séparent du nœud racine et la distance N qui sépare le concept subsument (CS) du nœud racine.

$$Sim_{Wu}(X,Y) = \frac{2N}{N1+N2}$$

$P(c_i \mid c_g)$ est donnée par le rapport entre la proximité sémantique qu'entretient le couple de concepts (c_g, c_i) et tous les couples (c_i, c_k).

$$P(c_i/c_g) = \frac{Sim_{Wu}(c_g,c_i)}{\sum_{c_k \in E} Sim_{Wu}(c_i,c_k)} \quad [12]$$

[5] Nous avons posé la contrainte que les concepts d'expansion c_g et c_s n'appartiennent pas à Q, pour ne pas tenir compte deux fois de ces concepts.

De la même manière, on estime la probabilité $P(c_i \mid c_s)$

Estimation des paramètres λ, θ et α

L'estimation automatique des paramètres λ, θ et α demeure un véritable défi. L'algorithme de maximisation de l'espérance (EM) est le plus utilisé. Dans ce travail, nous nous sommes contentés de les fixer aléatoirement (Voir exemple).

Après le remplacement de [11] dans [7] et par la suite de [7] dans [6], le modèle proposé est alors donné par la formule [13]:

$$P(Q \mid D) = \prod_{c_i \in Q} \left[(1-\lambda) \left[\begin{array}{c} \lambda P_{\overline{exp}}(c_i/D) + \\ \theta P_p(c_i/D) + (1-\theta) \\ \left[\alpha \sum_{\substack{c_g \in E \\ c_g \notin Q}} P(c_i \mid c_g)\, P(c_g \mid D) \;\; + (1-\alpha) \sum_{\substack{c_s \in E \\ c_s \in Q}} P(c_i \mid c_s)\, P(c_s \mid D) \right] \end{array} \right] \right] \quad [13]$$

4. Exemple

Dans cette section, nous allons présenter un exemple sur lequel nous appliquons le modèle proposé et le modèle uni-gramme. Une comparaison entre les résultats obtenus par les deux modèles est également présentée.

Supposons que l'on dispose d'une collection de deux documents D_1 et D_2 dont les indexes en mots simples sont:
D1={Natural (4), Science(6), Geology(10), Geography(5), Geophysics(8), Globe(3), aeroelastic (10) }
D2={Earth (7), Natural (3), Science (6), Anatomy (4), Regional (5)}
Soit la requête : Q= {earth, science, geography, aeroelastic}

4.1. Application du modèle uni-gramme

En fixant le paramètre de lissage λ= 0.5, l'application du modèle mixte uni-gramme de mots simples (uni-grammes lissé par interpolation linéaire) nous donne le résultat récapitulé dans le tableau suivant :

Score de Pertinence	Valeur
P(Q | D_2)	0.00225
P(Q | D_1)	0.00125

Tableau 1. *Résultat du modèle uni-gramme de mots simples*

Résultat : D2 est *plus pertinent* que D1 par rapport à Q.

4.2. Application du modèle à base de concepts proposé

Avant de passer à l'estimation des scores de pertinence pour chaque document, on génère les représentations conceptuelles de chaque document et de la requête. Les concepts correspondant à des entrées de l'ontologie sont représentés par les arborescences illustrées par la figure 1.

D1 = {Natural Science (9), Geology (10), Geography (5), Geophysics (8), Globe (3), aeroelastic (10)}
D2={Earth (7), Natural Science (7.5), Regional anatomy (8.5)}
Q= {earth science, geography, aeroelastic}

Il est important de remarquer que dans le modèle proposé, "aerolastic" est considéré comme un concept alors qu'il n'appartient pas à l'ontologie.

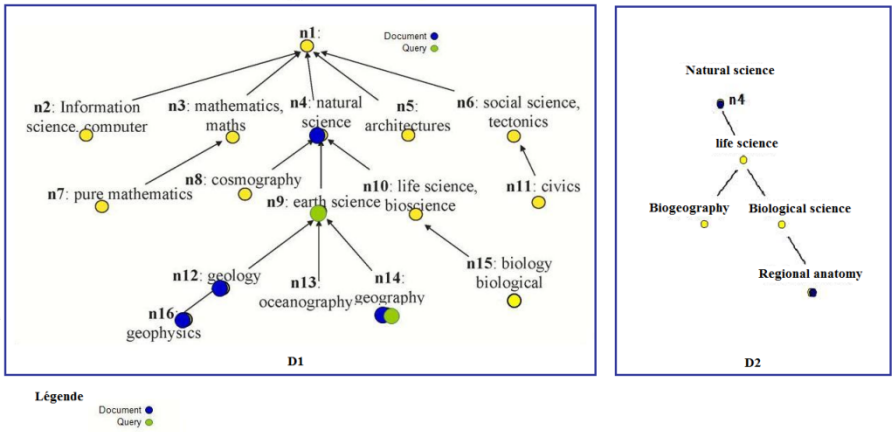

Figure 1. *Arborescences des concepts des documents D1 et D2*

Comme il est montré dans la figure 1, la projection de la requête sur les représentations arborescentes des deux documents nous permet de distinguer les différents niveaux de concepts que voici :

$c_{\overline{exp}}$ ={aeroelastic}
C_p= {earth science, geography}
C_g (earth science)={natural science, science, }
C_s (earth science)={geology, geography, oceanography, geophysics}
C_g (geography) = { }

Après l'expansion des concepts de la requête, nous pouvons désormais appliquer notre modèle. Il convient de souligner que nous fixons les paramètres (λ, θ, α) = (0.3, 0.5, 0.5). Ainsi, nous donnons un poids plus important (θ, α) au modèle d'expansion car c'est lui qui capture effectivement la sémantique de la requête et du document.

L'application numérique nous renvoie alors le résultat récapitulé dans le tableau 2.

Score de Pertinence	Valeur
$P(Q \mid D_1)$	0,000003726
$P(Q \mid D_2)$	0,000001481040

Tableau 2. *Résultat du modèle à base de concepts*

Résultat : D_1 est plus pertinent que D_2 par rapport à Q

Ce résultat est différent du premier (modèle uni-gramme). La fréquence élevée de "Earth" dans D2 a beaucoup influencé le premier résultat, alors qu'il est clair que, sémantiquement, D1 est plus proche de Q que D2. Ce constat a été bien mis en évidence par le modèle que nous proposons.

A travers cet exemple nous avons montré l'importance de la prise en compte des concepts et de leurs liens sémantiques sans pour autant négliger les mots ou les concepts qui n'existent pas dans l'ontologie.

5. Conclusion

Les modèles de langue se sont distingués par leur fondement sur la théorie des probabilités, ce qui leur procure une capacité à traiter de grandes masses de données ainsi qu'une flexibilité importante à intégrer diverses sources de connaissances (sémantiques). Cependant, l'hypothèse d'indépendance des mots pose un problème car elle dérive de la réalité du langage naturel et du coup, les requêtes et les documents sont représentés comme des sacs de mots. Pour apporter une solution à ce problème, nous proposons un nouveau modèle de langue qui s'inscrit dans la recherche conceptuelle d'information. Il est basé sur les concepts permettant ainsi d'intégrer de fortes dépendances entre les mots dans la requête (et dans les documents). De plus, en utilisant le lissage par interpolation linéaire, le modèle combine les concepts identifiés dans l'ontologie et ceux qui ne le sont pas (noms propres et néologismes). Ces derniers sont détectés comme des groupes de mots (de façon ad hoc) sélectionnés selon leurs poids d'apparition. Ce modèle est également un modèle d'expansion, les concepts de la requête, correspondant à des entrées de l'ontologie sont étendus par les concepts qui leur sont liés avec la relation de subsomption. Parmi ces derniers, on distingue les concepts de niveau spécifique et les concepts de niveau générique. Cette séparation est due à la différence de l'influence de ces deux niveaux sur les performances d'un SRI, alors que le premier niveau améliore le rappel, le second améliore la précision. Nous tenons compte de ce fait en utilisant le modèle de translation statistique lissé par interpolation linéaire. Le modèle proposé doit faire l'objet d'une évaluation sur une collection de référence (TREC) et sera comparé au modèle uni-gramme. Les résultats de cette évaluation nous permettront dans un premier temps de valider nos hypothèses sur la combinaison de plusieurs sources de concepts et dans un second temps de consolider notre approche en intégrant dans le modèle d'autres relations sémantiques présentes dans WordNet.

6. Références bibliographiques

Carmen Alvarez, Philippe Langlais et Jian-Yun Nie, « Mots composés dans les modèles de langue pour la recherche d'information », *TALN 2004 Session Poster Fès, 2004, p. 19–21.*

Shenghua Bao, Lei Zhang, Erdong Chen, Min Long, Rui Li, & Yong Yu, « LSM: Language Sense Model for Information Retrieval», *WAIM* , 2006, p. 97–108

Bai, J., Song, D., Bruza, P., Nie, J. Y. & Cao, G., « Query expansion using term relationships in language models for information retrieval», *CIKM*, 2005, p. 688-695.

M. Baziz, Indexation Conceptuelle Guidée par Ontologie pour la Recherche d'Information, thèse de doctorat, université de Paul Sabatier, 2005.

Berger, A. & Lafferty, J., « Information retrieval as statistical translation », *In Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, p. 222-229.

M. Boughanem, Mustapha Baziz & Henri Prade, « An Information Retrieval Driven by Ontology: from Query to Document Expansion», *RIAO*, 2007, p. 335–350.

Guihong Cao, Jian-Yun Nie & Jing Bai. « Integrating Word Relationships into Language Models », *SIGIR'05*, Salvador, Brazil, 2005.

Jianfeng Gao, Jian-Yun Nie, Guangyuan Wu & Guihong Cao, « Dependance Language model for information retrieval », *SIGIR'04*, Sheffield South Yorkshire, UK, 25-29, July 2004.

D. Lin, « An information-theoretic definition of similarity », *In Proceedings of 15th International Conference On Machine Learning*, 1998.

Jay M. Ponte & W. Bruce Croft, « A Language Modeling Approach to Information Retrieval », *Research and Development in Information Retrieval, Proc. ACM-SIGIR*, 1998, p. 275-281.

Song, F. & Croft, B, « A general language model for information retrieval» *In Proc. Of CIKM'99*, 1999, p. 316–321.

Stanley F. Chen & Joshua Goodman. «An Empirical Study of Smoothing Techniques for Language Modeling », Computer Science Group, Harvard University. Cambridge, Massachusetts, 1998.

Munirathnam Srikanth & Rohini Srihari, « Biterm Language Models for Document Retrieval » ACM *SIGIR'02,* Tampere, Finland. 2002

Munirathnam Srikanth & Rohini Srihari, « Incorporating Query Term Dependencies in Language Models for Document Retrieval » In *SIGIR'03,* Toronto, Canada, July 28–August 1, 2003.

Wu Z. & Palmer M, « Verb Semantics and Lexical Selection », *In Proceeding of the 32nd Annual Meeting of the Associations for Computational Linguistics*, 1994, pp. 133-138.

Zhai C & Lafferty J, « A Study of Smoothing Methods for Language Models Applied to Information Retrieval », *In Proceedings of the 2001 ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001, p. 334-342.