Modèle de Langue à base de Concepts pour la Recherche d'Information

Lynda SAID L'HADJ*, Mohand BOUGHANEM**

*Ecole Doctorale STIC, Ecole nationale Supérieure d'Informatique ESI, Algérie
l_said_lhadj@esi.dz
**Laboratoire IRIT, Université Paul Sabatier
118 route de Narbonne 31062 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. Plus précisément, ces modèles sont estimés à partir des mots simples apparaissant dans les documents 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, 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 d'intégrer les dépendances entre les mots en représentant les documents et les requêtes par les concepts.

1 Introduction

Les modèles de langue ont acquis une grande popularité en Recherche d'Information (RI) étant donné la solidité de leur fondement mathématique Ponte et Croft (1998). Ces modèles ne modélisent pas directement la notion de pertinence. Cette dernière est vue comme la probabilité conditionnelle ($P(Q|D)$) que la requête Q soit générée par le modèle de langue du document (D). $P(Q|D)$ est estimée sous l'hypothèse d'*indépendance* des mots qui simplifie le calcul mathématique. Cependant, elle pose un problème majeur lié à la représentation des documents (requêtes) comme des sacs de mots dénués de sémantique.

Pour pallier ce problème, une nouvelle génération de modèles de langue qui s'inscrit à l'intersection des modèles de langue et de la recherche sémantique d'information a été développée Cao et al. (2005), Srikanth et Srihari (2002, 2003). Dans cette intersection, deux grandes approches peuvent être distinguées : *l'approche statistique surfacique* qui prend en compte des dépendances surfaciques entre les mots et *l'approche sémantique* basée sur 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 renvoient ces mots. Même si l'approche conceptuelle souffre du problème de silence[1], nous pensons que la définition d'un modèle de langue mixte combinant les concepts identifiés dans l'ontologie et les concepts non identifiés permet de résoudre ce problème.

[1] dû à la non disponibilité de ressources conceptuelles complètes et générales.

Le reste du papier est organisé comme suit : Nous présentons un bref état de l'art sur les modèles de langue sémantiques dans la section 2. La section 3 est consacrée à la présentation du modèle proposé. Puis, nous déroulons un exemple dans la section 4. Enfin, dans la section 5, nous terminons ce papier avec une synthèse du travail présenté.

2 Etat de l'art

La pertinence d'un document face à une requête est en rapport avec la probabilité que la requête Q, vue comme une suite de mots $t_1t_2...t_n$, puisse être générée par le modèle de langue du document. Le score de pertinence est alors donné par :

$Score(Q,D)=P(Q|M_D)$[2]$=P(t_1t_2...t_n \mid M_D)$ (1)

Pour estimer la probabilité de (1), il faut que les t_i soient indépendants, ainsi :

$$P(Q|M_D) = \prod_{i=1}^{n} P(^{t_i}/_{D}) \quad (2)$$

L'hypothèse d'indépendance des mots pose deux problèmes majeurs : le premier est celui des *données éparses*, c'est-à-dire, si un mot t_i est absent dans le document, $P(Q \mid D)$ est alors nulle même si les autres $t_{j,j\neq i}$ sont présents. Ce problème a été résolu par les techniques de lissage[3] Zhai et al. (2001). Le second problème est la *représentation en sac de mots* qui ne permet pas la prise en compte de deux phénomènes très importants en RI à savoir la *polysémie* et la *synonymie*. Pour le résoudre, la plupart des travaux proposés jusque là a utilisé les techniques de lissage (le lissage sémantique) afin d'incorporer les sens des mots ainsi que les liens entre ces mots dans les modèles de langue. Ces travaux sont classés en deux catégories d'approches Cao et al. (2005): *L'approche statistique ou surfacique* et *l'approche guidée par les ressources sémantiques.*

L'approche surfacique tente d'intégrer les relations entre les mots selon des considérations statistiques, par exemple les cooccurrences. Le modèle de translation statistique de Berger et Lafferty (1999) fut l'un des premiers travaux dans cette direction. Gao et al. (2004), de leur coté, considèrent les dépendances entre les mots comme une variable cachée L, représentée par un graphe acyclique non orienté, pour modéliser les dépendances entre les mots. Srikanth et Srihari (2002) ont proposé un modèle bi-termes où ils ignorent la contrainte d'adjacence et de l'ordre des mots imposée dans les modèles bi-grammes. Enfin, Srikanth et Srihari (2003) présentent un modèle uni-gramme de concepts (des séquences de mots) identifiés avec un parseur syntaxique.

L'approche guidée par les ressources sémantiques se base sur des liens sémantiques extraits de ressources sémantiques comme les ontologies. Cao et al. (2005) ont proposé une approche qui combine le modèle d'indépendance (uni-gramme) avec le modèle de dépendance des mots en exploitant les techniques de lissage. Ces dépendances sont de deux types: statistique (cooccurrence) et sémantique (relations entre mots simples de WordNet). Dans la même direction, Bao et al. (2006) ont proposé un modèle de langue uni-gramme lissé avec un modèle uni-gramme de sens de ces mots identifiés par un système de désambiguïsation basé sur WordNet.

Les résultats des deux approches sont meilleurs que le modèle de langue uni-gramme. Cependant, l'approche surfacique engendre beaucoup de bruit et donc elle nécessite un filtrage linguistique voire sémantique. Nous pensons également que l'intégration de relations

[2] On écrit $P(Q|D)$ pour représenter la probabilité $P(Q|M_D)$ où M_D est le modèle de document.

[3] Attribuer une probabilité non nulle aux mots de la requête absents dans le document.

(surfaciques ou sémantiques) entre mots simples ou la simple désambigüisation de ces mots ne suffit pas pour capturer le contenu sémantique implicite des documents et des requêtes. Nous pensons de ce fait qu'un concept, correspondant par exemple à une entrée d'une ontologie, est plus précis qu'un mot isolé ou un sens isolé.

3 Modèle proposé

Le modèle que nous proposons se base sur les concepts. Plus précisément, tout document (respectivement requête) est projeté sur une ontologie par exemple WordNet, Nous utilisons à cet effet l'algorithme de Baziz (2005)[4] pour la détection des concepts. Ainsi, les termes ayant une entrée dans l'ontologie sont pris comme éléments du document qui permettent la construction de l'arborescence du document (de la requête). Mais, contrairement à Baziz (2005), nous proposons de garder les termes non reconnus dans l'ontologie dans le descripteur, car il peut arriver que ces termes renvoient à des concepts importants (cas des noms propres ou des néologismes).

Nous considérons la requête (Document) comme des sacs de concepts. Ainsi: $Q= c_1c_2...c_n$. Le score de pertinence est donné par P(Q | D) estimée comme suit:

$$P(Q \mid D) = \prod_{c_i \in Q} P(c_i/D) \quad (4)$$

Nous distinguons deux cas : Si c_i ne correspond à aucune entrée de WordNet, l'appariement est strict. Sinon, on exploite la hiérarchie de la requête et du document pour chercher non seulement c_i mais aussi les concepts qui lui sont proches. Le lissage par interpolation linéaire permet de tenir compte de ce fait, ainsi :

$$P(c_i/D) = \lambda P_{\overline{exp}}(c_i/D) + (1-\lambda) P_{exp}(c_i/D) \quad (5)$$

$P_{\overline{exp}}(c_i/D)$: est la probabilité que c_i non candidat à l'expansion soit généré par D.

$P_{exp}(c_i/D)$: est la probabilité que le concept c_i identifié dans WordNet soit généré par *D*.

Estimation de $P_{exp}(c_i \mid D)$. Quand le concept c_i correspond à une entrée de l'ontologie, on peut non seulement retrouver les documents où il figure effectivement (*appariement direct*) mais aussi les documents où figurent les concepts qui lui sont liés par des relations sémantiques de l'ontologie (subsomption) (*appariement indirect*). Ces deux cas sont combinés par un lissage par interpolation linéaire.

$$P_{exp}(c_i/D) = \theta P_p(c_i/D) + (1-\theta) P_{\bar{p}}(c_i/D) \quad (6)$$

$P_p(c_i/D)$ est 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 D.

Estimation des probabilités $P_p(c_i/D)$ et $P_{\overline{exp}}(c_i/D)$. Elles sont estimées en utilisant la formule de pondération de concepts CF *(Concept Frequency)* proposée dans Baziz (2005).

$$P_p(c_i/D) = P(c_{pi}/D) = \frac{cf(c_{pi},D)}{\sum_{c_j \in D} cf(c_j,D)} \quad (7)$$

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

Où *Count (c_{pi}, D)* retourne la fréquence d'apparition de c_{pi}, *Length(c_{pi})* représente le nombre de mots dans le concept c_{pi} et *sub_concept (c_{pi})* le nombre de tous les sous-concepts

[4] Cet algorithme comprend détection des groupes de mots, désambigüisation et pondération.

(des concepts de l'ontologie) dérivés de c_{pi}. Quand c_i ne correspond à aucune entrée de WordNet, nous annulons le deuxième terme de $cf(c_{pi}, D)$.
$cf(c_{\overline{exp}_i}, D) = Count(c_{\overline{exp}_i}, D)$

Quand ci (identifié ou non dans WordNet) est absent dans le document, alors P(Q | D) est nulle. C'est pourquoi nous lissons $P_p(c_i \mid D)$ et $P_{\overline{exp}}(c_i \mid D)$ en utilisant la méthode "Absolute Discount" appliquée aux concepts Zhai et al. (2001).

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

Où $|D| = \sum_{c_j \in D} cf(c_j, D)$ et $P_{MLE}(c_i \mid C) = \frac{cf(c_i,C)}{\sum_{c_i \in C} cf(c_i,C)}$ (C est le modèle de la collection).

Estimation de $P_{\bar{p}}(c_i \mid D)$. Pour intégrer les relations entre les concepts, nous utilisons le modèle de Berger et Lafferty (1999).

$P_{\bar{p}}(c_i \mid D) = \sum_{c_{\bar{p}} \in E} P(c_i \mid c_{\bar{p}})\, P(c_{\bar{p}} \mid D)$[5] (8)

Où E : est l'ensemble de tous les concepts $c_{\bar{p}}$ liés à ceux de la requête.

Les approches conceptuelles proposées jusque là, mélangent les concepts spécifiques avec les concepts génériques. Or, il a été constaté que les concepts génériques améliorent le rappel et que les concepts spécifiques améliorent la précision Baziz (2005), Zakos, J. (2005). Le modèle (8) est séparé pour tenir compte de ce fait par le lissage par interpolation linéaire:

$$P_{\bar{p}}(c_i \mid D) = \alpha \left[\sum_{\substack{c_g \in E \\ c_g \notin Q}} P(c_i \mid c_g)\, P(c_g \mid D) \right] + (1-\alpha) \left[\sum_{\substack{c_s \in E \\ c_s \notin Q}} P(c_i \mid c_s)\, P(c_s \mid D) \right] \quad (9)$$

Où : $P(c_i \mid c_g)$ (respectivement $P(c_i \mid c_s)$) est la probabilité conditionnelle que c_i soit généré par un concept plus générique c_g (respectivement c_s). Nous avons posé les contraintes $c_g \notin Q$ et $c_s \notin Q$ pour ne tenir qu'une seule fois de c_g et c_s quand ils existent déjà dans la requête.

Estimation de $P(c_i \mid c_g)$ et de $P(c_i \mid c_s)$. Ces probabilités interprètent la distance sémantique entre c_i et c_g (ou c_s). Elles sont estimées en utilisant une mesure de similarité basée sur la distance sémantique entre ces concepts. Nous avons choisi la mesure de Wu et Palmer (1994) car elle est simple à mettre en œuvre, de plus, elle est basée sur le principe de la distance sémantique 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 subsumant (CS) X et Y du nœud racine.

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

$P(c_i/c_g) = \frac{Sim_{Wu}(c_g,c_i)}{\sum_{c_k \in E} Sim_{Wu}(c_i,c_k)}$ (10)

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

Après remplacement des différentes probabilités dans [5], le modèle proposé est donné par :

$$P(Q \mid D) = \prod_{c_i \in Q} \left[(1-\lambda) \left[\begin{array}{c} \lambda P_{\overline{exp}}(c_i/D) + \\ \theta P_p(c_i/D) + (1-\theta) \\ \left[\alpha \sum_{\substack{c_g \in E \\ c_g \notin Q}} P(c_i \mid c_g)\, P(c_g \mid D) + (1-\alpha) \sum_{\substack{c_s \in E \\ c_s \notin Q}} P(c_i \mid c_s)\, P(c_s \mid D) \right] \end{array} \right] \right] \quad (11)$$

[5] Il s'agit du modèle de l'expansion des concepts à proprement parler

4 Exemple

Soient: 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)}
Q= {earth, science, geography, aeroelastic}

4.1 Application du modèle uni-gramme mixte

Le modèle de langue mixte est donné par : $P(Q|D) = \prod_{i=1}^{K}[\lambda P(t_i|D) + (1-\lambda)P(t_i|C)]$. Sachant que les $t_{i..k}$ correspondent aux mots de la requête et λ est fixé à 0.5, l'application numérique du modèle retourne les scores de pertinences suivants : $P(Q \mid D_2) = 0.00225$ et $P(Q \mid D_1) = 0.00125$. On remarque ainsi que D2 est plus pertinent que D1 par rapport à Q.

4.2 Application du modèle à base de concepts proposé

La projection de D1 et de D2 sur la hiérarchie de WordNet retourne les représentations conceptuelles : 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}

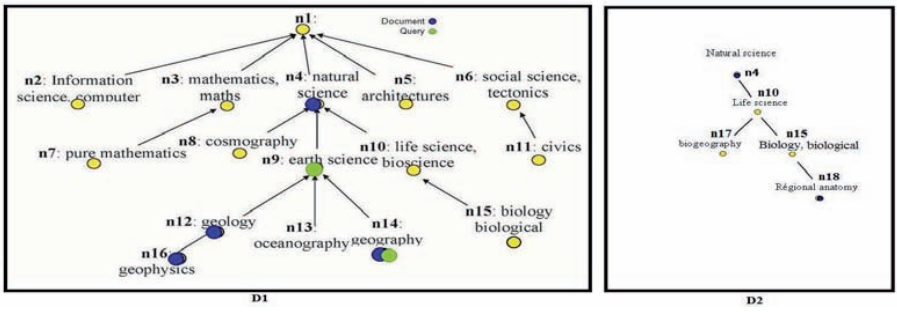

FIG. 1-Représentations hiérarchiques des concepts de D1 et D2.

La figure 1 permet de distinguer les 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, geology, geophysics}, C_g (geography) = { }.

Après application du modèle proposé[6], les scores de pertinence sont : $P(Q \mid D_1) = 0{,}000003726$ et $P(Q \mid D_2) = 0{,}000001481040$. Nous remarquons alors que D_1 est *plus pertinent* que D_2 par rapport à Q. Ce résultat diffère du premier. Il est justifié par la fréquence élevée de « Earth » de D2 (dont le sens est loin de celui de la requête) qui a influencé le résultat retourné par le modèle uni-gramme.

[6] Les paramètres (λ, θ, α) = (0.3, 0.5, 0.5). Nous avons donné un poids plus important (0.5) au modèle d'expansion car il capture effectivement la sémantique de la requête et du document.

5 Conclusion

Dans ce papier nous avons proposé un modèle de langue basé sur les concepts. Le choix d'intégrer les concepts dans les modèles de langue se justifie aussi bien par les résultats prometteurs de la recherche conceptuelle d'information que par la performance et la flexibilité des modèles de langue à intégrer plusieurs sources de connaissances. En effet, cette flexibilité nous a permis de tenir compte non seulement des concepts de l'ontologie ainsi que de leurs liens mais aussi des concepts qui n'apparaissent pas dans l'ontologie. Ce modèle est en cours d'expérimentation. Les résultats nous permettront d'approfondir et de consolider nos hypothèses sur la combinaison mots simples-concepts ainsi que de la séparation des deux niveaux de concepts génériques et concepts spécifiques.

6 Références bibliographiques

Shenghua Bao, Lei Zhang, Erdong Chen, Min Long, Rui Li, and Yong Yu (2006). *LSM: Language Sense Model for Information Retrieval*, WAIM, pp. 97–108.

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

Berger, A. and Lafferty, J., (1999). *Information retrieval as statistical translation*, In Proc. of the 1999 ACM SIGIR, pp. 222-229.

Guihong Cao, Jian-Yun Nie, Jing Bai (2005). *Integrating Word Relationships into Language Models*,. SIGIR'05, Salvador, Brazil, August 15–19.

Jianfeng Gao, Jian-Yun Nie, Guangyuan Wu, Guihong Cao (2004). *Dependance Language model for information retrieval*, SIGIR'04.

Jay M. Ponte and W. Bruce Croft (1998). *A Language Modeling Approach to Information Retrieval*, Proc. of ACM-SIGIR, pp. 275-281.

Munirathnam Srikanth et Rohini Srihari (2002). *Biterm Language Models for Document Retrieval*, ACM SIGIR'02, Tampere, Finland.

Munirathnam Srikanth, Rohini Srihari (2003). *Incorporating Query Term Dependencies in Language Models for Document Retrieval*, In SIGIR'03, Canada.

John Zakos (2005). *A Novel Concept and Context based Approach for Web Information Retrieval*, Doctorate thesis, Griffith University, 2005.

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

Zhai C and Lafferty J (2001). *A Study of Smoothing Methods for Language Models Applied to Information Retrieval*, In Proc. of the 2001 ACM SIGIR. pp. 334-342.