TALN 2004, Session Poster, Fès, 19–21 avril 2004

Mots composés dans les modèles de langue pour la recherche d'information

Carmen Alvarez, Philippe Langlais et Jian-Yun Nie
RALI/IRO, Université de Montréal
CP. 6128, succursale Centre-ville
Montréal, Québec, H3C 3J7 Canada
{bissettc,felipe,nie}@iro.umontreal.ca

Résumé - Abstract

Une approche classique en recherche d'information (RI) consiste à bâtir une représentation des documents et des requêtes basée sur les mots simples les constituant. L'utilisation de modèles bigrammes a été étudiée, mais les contraintes sur l'ordre et l'adjacence des mots dans ces travaux ne sont pas toujours justifiées pour la recherche d'information. Nous proposons une nouvelle approche basée sur les modèles de langue qui incorporent des *affinités lexicales* (ALs), c'est à dire des paires non ordonnées de mots qui se trouvent proches dans un texte. Nous décrivons ce modèle et le comparons aux plus traditionnels modèles unigrammes et bigrammes ainsi qu'au modèle vectoriel.

Previous language modeling approaches to information retrieval have focused primarily on single terms. The use of bigram models has been studied, but the restriction on word order and adjacency may not be justified for information retrieval. We propose a new language modeling approach to information retrieval that incorporates lexical affinities (LAs), or pairs of words that occur near each other, without a constraint on word order. We explore the use of LAs in a language modeling approach, and compare our results with the vector space model, and unigram and bigram language model approaches.

Mots-clefs – Keywords

Modèles de langue, recherche d'information, mots composés
Language models, information retrieval, compound terms, word pairs

1 Introduction

L'utilisation des modèles de langue en RI a été introduite par Ponte et Croft (1998). Chaque document est considéré comme un échantillon d'un langage particulier, et un modèle de langue est entraîné pour chaque document. Pour une requête donnée, les documents sont triés par ordre décroissant de la probabilité que le modèle du document génère la requête. Cette approche donne des performances comparables, voire supérieures, au modèle vectoriel.

Représenter un document par un modèle de langue représente un certain nombre de désavantages dont le principal est le problème ici aigu de la sous-représentation des données d'entraînement. Entraîner un modèle de langue (même un simple unigramme) sur des documents qui contiennent quelques centaines de mots représente en effet un certain défi. Ainsi, Song et Croft (1999) étudient différentes techniques de lissage connues comme le lissage Good-Turing ou la combinaison linéaire de plusieurs modèles n-grammes d'ordres différents. Hiemstra (2002) propose une technique d'interpolation où un modèle unigramme du document et un modèle de corpus sont combinés. Lavrenko et Croft (2001) font également usage d'une combinaison de modèles de documents et d'un modèle de corpus pour estimer un modèle de pertinence (probabilité qu'un mot soit pertinent pour une requête), sans nécessiter de données d'entraînement spécifiques.

Nous proposons une approche basée sur l'entraînement de modèles de langues qui incorporent des paires de mots non définies par des contraintes d'adjacence ou d'ordonnancement: les affinités lexicales (ALs). Nous commençons par présenter en section 2 le modèle unigramme que nous utilisons à des fins comparatives dans nos expériences. Nous décrivons ensuite en section 3 une procédure initialement proposée par Maarek *et al.* (1991) qui permet d'obtenir les ALs d'un document et présentons en section 4 un modèle de langue faisant usage de ces affinités. Nous décrivons ensuite en section 5 le cadre expérimental qui nous a permis d'étudier le comportement des différents modèles décrits et discutons nos résultats dans la section 6. Nous montrons en particulier qu'un modèle de langue unigramme lissé rivalise avec l'approche classique du modèle vectoriel et que la prise en compte des affinités lexicales améliore de manière sensible les performances.

2 Modèles n-gramme sur les mots simples

Le score de pertinence d'un document d pour une requête de N mots $q = w_1^N = w_1, \ldots, w_N$ est donné par la probabilité que le modèle du document génère la requête. Dans le cas d'un modèle n-gramme p_{n_d}, cette pertinence s'exprime simplement par l'équation 1 où n représente l'ordre du modèle:

$$score(d, q) = \prod_{i=1}^{N} p_{n_d}(w_i | w_{i-n+1}^{i-1}) \tag{1}$$

Réaliser un système de RI à l'aide de modèles de langue peut se résumer dans sa forme la plus simple à entraîner autant de modèles que de documents. La probabilité qu'un mot w d'une requête soit généré par le modèle de langue d'un document peut alors être estimée par le maximum de vraisemblance (MLE), ce qui revient dans le cas unigramme à calculer la fréquence relative de w dans le document d.

L'estimateur à maximum de vraisemblance directement injecté dans l'équation 1 s'avère en pratique très peu utile: les documents qui ne contiennent pas l'ensemble des mots de la requête se voient attribuer un score de pertinence nul. Ce problème bien connu en modélisation de la langue (le lissage) est ici particulièrement épineux puisque les documents que nous traitons contiennent environ 200 à 400 mots.

La problématique du lissage a été et continue à être un objet d'investigation scientifique et de nombreuses techniques ont été proposées pour l'entraînement de modèles de langue à partir de grands corpus de textes (Goodman, 2001). Nous étudions dans Alvarez *et al.* (2003) différentes techniques de lissage spécifiques à l'entraînement de modèles de langue pour la RI et rapportons ici les configurations pour lesquelles nous avons observé les meilleurs résultats. Il convient de noter que les documents ainsi que les requêtes sont soumis à un pré-traitement qui consiste en une lemmatisation et en la suppression de mots apparaissant dans une *stopliste*[1].

Dans le cas d'un modèle de langue unigramme, nous combinons linéairement le modèle MLE avec un modèle de corpus selon l'équation 2. Ce dernier est un modèle unigramme MLE sur l'ensemble des documents de la collection. Dans le cas d'un modèle bigramme, nous combinons linéairement le modèle MLE du document avec le modèle unigramme selon l'équation 3:

$$p_{uni_d}(w) = \lambda_1 p_{MLE_d}(w) + (1 - \lambda_1) p_{corpus}(w) \tag{2}$$

$$p_{bi_d}(w_i|w_{i-1}) = \lambda_2 p_{MLE_d}(w_i|w_{i-1}) + (1 - \lambda_2) p_{uni_d}(w_i) \tag{3}$$

3 Affinités lexicales

L'hypothèse d'indépendance entre mots, faite par le modèle unigramme, ainsi qu'une grande partie des approches à la RI, n'est pas toujours justifiée. Les modèles bigramme (et à fortiori les modèles d'ordre supérieur) tentent en effet de rendre compte des dépendances entre termes; tout en supposant que l'ordre des mots est important. Tandis que cette dernière hypothèse semble raisonnable pour des applications comme la reconnaissance de parole, elle ne s'applique pas nécessairement à la RI. Par exemple, pour une requête "apartment rentals", un document contenant les termes "rent an apartment" ne doit pas être à priori moins bien classé qu'un autre document contant les termes "apartments for rent". Notre réponse à ce problème consiste à baser notre modélisation sur une unité lexicale n'imposant aucune restriction sur l'ordre de ses mots et peu de contrainte sur leur adjacence: l'affinité lexicale.

Selon Martin *et al.* (1983), 98% des relations lexicales dans un texte mettent en jeu des mots dans une fenêtre de 5 mots. Nous adoptons cette propriété pour identifier les unités (paires de mots) sur lesquelles bâtir nos modèles de langue. Par ailleurs, Maarek *et al.* (1991) introduisent le concept de *pouvoir de résolution* d'une paire de mots. À l'instar des facteurs *tf* et *idf* utilisés dans le modèle vectoriel, l'idée principale derrière le pouvoir de résolution est que les paires de mots qui caractérisent le mieux un document sont celles qui ont en même temps une fréquence élevée dans le document et une fréquence relativement basse dans la collection. Les auteurs suggèrent de calculer le pouvoir de résolution d'une paire $< u, v >$ pour un document d selon l'équation 4; où $c_d(< u, v >)$ est la fréquence de la paire dans le document d. Le terme logarith-

[1] Une liste de 571 mots anglais fournie avec le système SMART a été utilisée.

mique dans cette équation peut être vu comme une approximation de la quantité d'information véhiculée par la paire, comparable au facteur *idf*.

$$\rho_d(< u, v >) = -c_d(< u, v >) \times \log(p_{corpus}(u) \times p_{corpus}(v)) \tag{4}$$

Le pouvoir de résolution de toutes les paires de mots distants d'au plus cinq mots (pleins) dans un document est calculé. Il est important de noter que les paires $< u, v >$ sont stockées par ordre lexicographique (la paire "traduction automatique" vue dans un texte est traitée comme "automatique,traduction"). La table 1 montre les cinq meilleurs affinités lexicales de deux documents de notre collection.

AP900302-10			AP900427-3		
AL ($< u, v >$)	fréquence	ρ_d	AL ($< u, v >$)	fréquence	ρ_d
court supreme	7	43,6	union violence	6	37,7
court property	6	38,3	greyhound violence	5	37,5
public sidewalk	4	30,5	greyhound union	5	34,6
court night	5	29,7	member union	6	34,2
court justice	4	24,6	condone violence	4	33,4

Table 1: Les 5 meilleures ALs selon le pouvoir de résolution pour deux documents de la collection TREC AP90: le document AP900302-10 traitant du 200è anniversaire de la cour suprême des États-Unis et le document AP900427-3 traitant des syndicats de la société Greyhound.

4 M_{LA}: Un modèle de langue basé sur les affinités lexicales

Notre modèle M_{LA} est un modèle unigramme qui estime les probabilités des mots simples et des paires de termes du document. Pour ce faire, on introduit les comptes décrits en équation 5; desquels on obtient la probabilité de chaque événement $p(w')$ (w' étant un mot ou une paire) en les normalisant par la constante $D = \sum_{w \in d} c_d(w) + \sum_{< u,v > \in d} \beta \rho_d(< u, v >)$. Cette approche est conceptuellement équivalente à l'ajout dans chaque document du compte fractionnaire (contrôlé par β_d, fixe pour l'ensemble des documents) des affinités lexicales du document.

$$c^\star_d(w') = \begin{cases} c_d(w) & \text{si } w' \text{ est un mot simple } w \\ \beta_d \rho_d(< u, v >) & \text{sinon} \end{cases} \tag{5}$$

Ce modèle est à son tour lissé par le modèle de corpus décrit dans la section 2, où les probabilités $p_{corpus}(w')$ se basent sur les comptes $c^\star_{corpus}(w')$, contrôlés par un facteur β_{corpus}.

$$p_{M_{LA}}(w') = \lambda_{M_{LA}} \frac{c^\star(w')}{D} + (1 - \lambda_{M_{LA}}) p_{corpus}(w')) \tag{6}$$

Si le même traitement est appliqué à la requête q (en contrôlant les comptes fractionnaires par un coefficient β_q, fixe pour toutes les requêtes), le score de pertinence est alors:

$$score(d, q) = \prod_{w' \in q} p_{M_{LA_d}}(w')^{c^\star_q(w')} \tag{7}$$

5 Expériences

Nous avons étudié le comportement des différents modèles présentés sur la collection TREC AP90, qui contient 78 321 documents en anglais de l'Associated Press newswire de 1990. 53 requêtes étiquetées manuellement pour les pistes translinguistiques des campagnes TREC-6 et TREC-7 constituaient notre corpus de test (une moyenne de 22 documents pertinents sont associés à chaque requête). Chaque requête comporte un champ titre de 1 à 5 mots (2.5 en moyenne) ainsi qu'un champ description contenant de 3 à 19 mots (7 en moyenne).

Chaque expérience comprend deux tests. L'un dénoté TITRE consiste à n'utiliser que le champ titre d'une requête, l'autre dénoté DESC utilise les champs titre et description. La tâche du système consiste à classer par ordre décroissant de pertinence les documents de la collection pour chaque requête. La métrique d'évaluation que nous utilisons est la *précision moyenne* habituellement utilisée dans ce type de tâche et qui mesure la moyenne de précision obtenue sur plusieurs points de rappel (Salton & McGill, 1983), pour les 1000 premiers documents retrouvés par le système. Afin de comparer nos différents modèles à un système éprouvé, nous avons utilisé le système SMART qui implémente le modèle vectoriel classique (Buckley, 1985).

La table 2 montre les performances du modèle M_{LA}. Les précisions indiquées en gras correspondent à des variantes dont les performances mesurées dépassent le modèle unigramme et le système SMART. Notre modèle dépasse le modèle unigramme pour plusieurs combinaisons de β_d et β_q. Pour les requêtes TITRE, le gain relatif mesuré le plus important est de 2.0%. La meilleure précision moyenne obtenue avec les requêtes DESC (43.20) est supérieure au modèle unigramme (42.75), soit un gain relatif de 1.0%.

	$\lambda_{M_{LA}}$ β_d,β_q,β_{corpus}	0.1	0.2	0.4	0.5	0.6	Smart	1-gram	2-gram
TITRE	.01,.01,.01	39.20	**40.77**	**40.87**	**41.43**	**41.13**	33.49	40.71	40.72
	.01,.01,.0001	39.39	**40.92**	**40.95**	**41.54**	**41.39**	33.49	40.71	40.72
DESC	.01,.005,.01	41.18	***43.20***	42.66	42.29	41.96	34.98	42.75	42.77
	.04,.005,.0001	40.95	***43.14***	42.74	42.39	42.00	34.98	42.75	42.77

Table 2: Précision moyenne obtenue par le modèle M_{LA} en fonction des méta-paramètres $\lambda_{M_{LA}}$ et β_d pour le document, β_q pour la requête et β_{corpus} pour le modèle de corpus. Les résultats qui sont significatifs, selon le test de Wilcoxon des rangs signés avec un intervalle de confiance de 95%, sont indiqués en italique.

6 Discussion

Il existe plusieurs travaux qui tentent d'améliorer les performances d'un système de recherche d'information basé sur les mots simples. En particulier, Nie et Dufort (2002) étudient l'ajout de termes composés comme de nouveaux indices dans le modèle vectoriel. Ils montrent que l'adjonction de termes en provenance des bases terminologiques (Termium[2] et la Banque de terminologie du Québec[3]) permet d'améliorer les performances d'une tâche de RI si ces termes

[2]http://www.termium.com

[3]http://www.olf.gouv.qc.ca/ressources/bibliotheque/dictionnaires/Internet/Index/

sont incorporés de manière adéquate.

Dans cette étude, nous montrons qu'il est possible d'améliorer (certes de manière modeste) les performances d'un modèle de langue s'il fait usage d'affinités lexicales. Ce type d'unité est intuitivement attirant en RI car il ne possède pas la rigidité des séquences de mots. Nous observons que l'augmentation en performance due aux ALs est plus marquée pour les requêtes courtes (TITRE). Les requêtes plus longues (DESC) contiennent en moyenne 18 paires. Bien que n'ayant pas mené d'analyse systématique, nous pensons que dans le cas de requêtes longues, plusieurs ALs ne sont pas pertinentes.

Les perspectives que cette étude suggère sont multiples. Les modèles que nous proposons sont régis par plusieurs paramètres (λ_1, λ_2, $\lambda_{M_{LA}}, \beta_d$) que nous avons fixés empiriquement. Une approche plus systématique nous permettrait d'ajuster ces paramètres et d'en augmenter le nombre. Il est en effet intuitif de penser que le poids donné à un modèle donné devrait à tout le moins être conditionné par la taille du document traité. Nous souhaitons également étudier l'impact de differentes techniques de filtrage des ALs, notamment grâce à un étiqueteur morpho-syntaxique.

Références

ALVAREZ C., LANGLAIS P. & J.Y-NIE (2003). *Word Pairs in Language Modeling for Information Retrieval*. Rapport interne, RALI.

BUCKLEY C. (1985). *Implementation of the SMART information retrieval system*. Rapport interne, Cornell University. Technical report 35-686.

GOODMAN J. (2001). A bit of progress in language modeling. *Computer Speech and Language*, p. 403–434.

HIEMSTRA D. (2002). Term-specific smoothing for the language modeling approach to information retrieval: the importance of a query term. In *25th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, p. 35–41, Tampere, Finland.

LAVRENKO V. & CROFT W. B. (2001). Relevance-based language models. In *24th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, p. 120–127.

MAAREK Y., BERRY D. & KAISER G. (1991). An information retrieval approach for automatically constructing software libraries. *IEEE transactions on software engineering*, p. 800–813.

MARTIN W., AL B. & VAN STERKENBURG P. (1983). On the processing of a text corpus: From textual data to lexicographical information. In E. R.R.K. HARTMANN, Ed., *Lexicography: Principles and Practice*, Applied Language Studies Series. Academic Press, London.

NIE J.-Y. & DUFORT J. (2002). Combining words and compound terms for monolingual and cross-language information retrieval. In *Information 2002*.

PONTE J. M. & CROFT W. B. (1998). A language modeling approach to information retrieval. In *21st annual international ACM SIGIR conference on Research and Development in Information Retrieval*, p. 275–281, Melbourne, Australia.

SALTON G. & MCGILL M. J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw Hill.

SONG F. & CROFT W. B. (1999). A general language model for information retrieval. In *22nd annual international ACM SIGIR conference on Research and Development in Information Retrieval*, p. 279–280.