Modèles d'information pour la recherche multilingue

Bo Li, Eric Gaussier

Université J. Fourier/Grenoble 1 - CNRS
Centre Equation 4, UFR IM2AG, LIG/AMA - BP 53 - F-38041 Grenoble Cedex 9
bo.li@imag.fr, eric.gaussier@imag.fr

RÉSUMÉ. *Nous présentons dans cet article plusieurs extensions multilingues des modèles d'information, en particulier le modèle log-logistique (LL) et le modèle Smoothed Power Law (SPL), récemment introduits en recherche d'information (Clinchant* et al., *2010). Ces extensions sont fondées sur (a) une généralisation de la notion d'information utilisée dans ces modèles, (b) une généralisation des variables aléatoires utilisées et (c) une expansion de la requête utilisant l'ensemble des traductions de chaque mot. Nous analysons ensuite ces extensions d'un point de vue théorique, par l'intermédiaire d'une condition que doivent satisfaire les systèmes de recherche d'information multilingue. Cette nouvelle condition permet d'étendre le champ de l'approche axiomatique de la recherche d'information au cadre multilingue. Les résultats expérimentaux, obtenus sur trois collections et trois couples de langue, sont en accord avec l'analyse théorique et montrent que le modèle LL fournit les meilleurs résultats en recherche d'information multilingue.*

ABSTRACT. *We present in this paper well-founded cross-language extensions of the recently introduced models in the information-based family for information retrieval, namely the LL (log-logistic) and SPL (smoothed power law) models of (Clinchant* et al., *2010). These extensions are based on (a) a generalization of the notion of information used in the information-based family, (b) a generalization of the random variables also used in this family, and (c) the direct expansion of query terms with their translations. We then review these extensions from a theoretical point-of-view, prior to assessing them experimentally. The results of the experimental comparisons between these extensions and existing CLIR systems, on three collections and three language pairs, reveal that the cross-language extension of the LL model provides a state-of-the-art CLIR system, yielding the best performance overall.*

MOTS-CLÉS : *Recherche d'information multilingue, modèles d'information*

KEYWORDS: *Cross-language information retrieval, information models*

CORIA 2012, pp. 9–24, Bordeaux, 21-23 mars 2012

1. Introduction

La recherche d'information multilingue consiste à interroger une collection de documents rédigés dans une (ou des) langue différente de celle utilisée pour formuler le besoin d'information. Si les tentatives pour prendre en compte une dimension multilingue en recherche d'information datent de la fin des années 60, un renouveau pour cette problématique est apparu dans les années 90 avec l'émergence du web et la disponibilité d'un grand nombre de pages écrites dans des langues différentes. Si les organisations internationales et les gouvernements de pays "multilingues", par exemple, ont toujours été des utilisateurs de systèmes de recherche d'information multilingue, le besoin de tels systèmes pour la vie de tous les jours se développe avec l'ensemble des activités liées au tourisme et au commerce électronique[1].

Il y a plusieurs façons de franchir la barrière de la langue dans un modèle de recherche d'information multilingue : en projetant la représentation des documents sur l'espace de la requête (c'est l'approche "traduction des documents"), en projetant la requête sur l'espace des documents ("traduction de la requête"), ou en projetant ces deux représentations dans un espace commun ("approche interlingue"). Une fois une stratégie choisie, le couplage entre traduction et recherche d'information se fait généralement soit par le biais d'approches indépendantes du modèle, soit par le biais d'approches dépendantes du modèle. Dans les approches indépendantes du modèle, le couplage entre traduction et recherche d'information est faible, les documents, ou les requêtes, étant tout d'abord traduits, et une simple recherche monolingue étant alors utilisée. Un exemple prototypique ici est celui des modèles fondés sur la traduction automatique (e.g. (Kraaij *et al.,* 2003, Braschler, 2004)) qui fait appel à un système de traduction automatique existant pour traduire documents ou requêtes. Les approches dépendantes du modèle intègrent traduction et recherche d'information dans un même cadre. Le couplage ici est donc fort. De telles approches, dont un exemple pour les modèles de langue est développé dans (Federico *et al.,* 2002, Kraaij *et al.,* 2003), ont l'avantage de permettre au processus de recherche d'information de tenir compte des incertitudes liées au processus de traduction.

La plupart des approches dépendantes du modèle sont des extensions, au cadre multilingue, de modèles de recherche d'information développés initialement dans un cadre monolingue. Si la majorité des modèles monolingues ont été étendus dans un cadre multilingue, cela n'est pas le cas de tous, et nous explorons dans cet article l'extension multilingue de la famille des modèles d'information, récemment introduite en recherche d'information. Deux modèles de cette famille ont fourni, dans un cadre monolingue, des résultats équivalents ou supérieurs aux autres modèles sur plusieurs collections, et la question de leur comportement dans un cadre multilingue reste ouverte. C'est précisément cette question que nous étudions ici.

---

1. Le livre récent de J.-Y. Nie sur la recherche d'information multilingue (Nie, 2010) expose en détail les besoins de systèmes de recherche d'information multilingue, et nous renvoyons les lecteurs intéressés par cet aspect à ce livre.

Notre article est structuré de la façon suivante : nous introduisons en section 2 la famille des modèles d'information, avant de présenter trois extensions multilingues possibles qui sont explicitées sur deux modèles de cette famille ; cette section introduit aussi une nouvelle condition, appelée condition D/C pour Dilution/concentration, sous la forme d'un critère formel que les modèles de recherche d'information multilingue doivent vérifier. La section 3 présente ensuite les résultats expérimentaux obtenus avec les modèles précédents, sur trois collections et trois couples de langue.Enfin, la section 4 conclut et article. Les notations que nous utilisons dans la suite sont résumées dans le tableau 1 (w représente un mot).

Tableau 1. *Notations utilisées dans l'article*

Notation	Description
x_w^q	Nombre d'occurrences de w dans la requête q
x_w^d	Nombre d'occurrences de w dans le document d
t_w^d	Version normalisée de x_w^d
l_d	Longueur du document d
l_m	Longueur moyenne d'un document
L	Longueur (nombre de *tokens*) de la collection
N	Nombre de documents dans la collection
$DS(w)$	Ensemble de documents contenant le mot w
N_w	Nombre de documents contenant le mot w ($N_w = \|DS(w)\|$)
$TS(w)$	Ensemble de traductions de w
RSV(q, d)	*Retrieval Status Value* (score) du document d pour la requête q

2. Modèles d'information

Les modèles d'information, récemment introduits dans (Clinchant *et al.,* 2010), calculent la similarité entre requêtes et documents à travers la quantité d'information apportée par les termes du documents sur la requête. Deux instances particulières de ces modèles, le modèle Log-Logistique (que nous dénoterons LL) et le modèle *Smoothed Power Law* (que nous dénoterons SPL), étudiées dans (Clinchant *et al.,* 2010, Clinchant *et al.,* 2011), produisent des résultats soit de même qualité soit de qualité supérieure que les autres modèles de recherche d'information, et ce sur plusieurs collections et différents cadres, comme celui de la rétropertinence. Nous étudions ici les extensions possibles de ces modèles à un cadre multilingue.

Les modèles d'information sont fondés sur le score suivant[2] :

$$\begin{aligned} RSV(q,d) &= \sum_{w \in q} -\frac{x_w^q}{l_q} \log P(X_w \geq t_w^d | \lambda_w) \\ &= \sum_{w \in q \cap d} -\frac{x_w^q}{l_q} \log P(X_w \geq t_w^d | \lambda_w) \end{aligned} \quad [1]$$

où :

– t_w^d est une fonction de normalisation dépendant du nombre d'occurrences, x_w^d, de w dans d et de la longueur, l_d, de d, et satisfait :

$$\frac{\partial t_w^d}{\partial x_w^d} > 0; \;\; \frac{\partial t_w^d}{\partial l_d} < 0; \frac{\partial^2 x_w^d}{\partial (t_w^d)^2} \geq 0$$

Dans ce travail, suivant en cela (Clinchant *et al.,* 2010), cette fonction est définie par : $t_w^d = x_w^d \log(1 + c\frac{l_m}{l_d})$, avec c un paramètre de normalisation ;

– P est une distribution de probabilité définie pour les variables aléatoires X_w associées à chaque mot w et ayant pour domaine les valeurs possibles de t_w^d. Cette distribution doit être :

- Continue, les valeurs prises par les variables aléatoires considérées étant continues ;

- Compatible avec le domaine de t_w^d, i.e. si t_{min} est la valeur minimale de t_w^d, alors $P(X_w \geq t_{\text{min}} | \lambda_w) = 1$;

- "En rafale", i.e. qu'elle doit satisfaire :
$\forall \epsilon > 0, \;\; g_\epsilon(x) = P(X \geq x + \epsilon | X \geq x)$ est strictement croissante en x ;

– Et λ_w est un paramètre dépendant de la collection. Comme suggéré dans (Clinchant *et al.,* 2010), il est fixé dans cette étude à :

$$\lambda_w = \frac{N_w}{N} \quad [2]$$

Comme on peut le remarquer, l'équation 1 calcule l'information apportée par le document sur chaque mot de la requête ($-\log P(X_w \geq t_w^d | \lambda_w)$) pondérée par l'importance du mot dans la requête ($\frac{x_w^q}{l_q}$). En choisissant pour P les distributions log-

2. Nous introduisons une légère modification, à savoir la normalisation par la longueur de la requête, dans la formule donnée dans (Clinchant *et al.,* 2010), et ce afin de présenter de façon plus intuitive ces modèles. Cette modification ne change pas l'ordre des documents.

logistique et SPL, on obtient les deux modèles LL et SPL définis dans (Clinchant *et al.,* 2010) :

$$RSV_{LL}(q,d) = \sum_{w \in q \cap d} -\frac{x_w^q}{l_q} \log \frac{\lambda_w}{\lambda_w + t_w^d}$$

$$RSV_{SPL}(q,d) = \sum_{w \in q \cap d} -\frac{x_w^q}{l_q} \log \frac{\lambda_w^{\frac{t_w^d}{t_w^d+1}} - \lambda_w}{1-\lambda_w}$$

Nous allons maintenant étudier les extensions multilingues de cette famille de modèles.

2.1. *Extensions multilingues*

L'information apportée par un document sur un terme de la requête dans l'équation 1 est restreinte à ce mot dans le document (les autres mots n'interviennent pas dans le calcul de cette information). Il est cependant possible d'adopter un point de vue plus général en considérant l'information moyenne apportée par tous les mots du document *reliées* au mot de la requête. Soit $\mathcal{F}(w)$ l'ensemble de tous les mots reliés, par le biais d'une relation que nous ne spécifions pas pour l'instant, au mot w. Soit de plus la *relation normalisée* entre w et un mot w' de d, une quantité notée $\mathcal{A}(w,w',d)$, définie par :

$$\mathcal{A}(w,w',d) = \begin{cases} \frac{I_{\mathcal{F}(w)}(w')}{\sum_{w'' \in d} I_{\mathcal{F}(w)}(w'')} & \text{if } \sum_{w'' \in d} I_{\mathcal{F}(w)}(w'') > 0 \\ 0 & \text{otherwise} \end{cases}$$

où $I_{\mathcal{F}(w)}$ est la fonction indicatrice de l'ensemble $\mathcal{F}(w)$. L'information moyenne apportée par tous les mots du document d reliés à un terme de la requête w donné peut alors être définie par : $-\sum_{w' \in d} \mathcal{A}(w,w',d) \log P(X_{w'} \geq t_{w'}^d | \lambda_{w'})$, ce qui conduit à la fonction de recherche suivante :

$$RSV(q,d) = -\sum_{w \in q} \frac{x_w^q}{l_q} \sum_{w' \in d} \mathcal{A}(w,w',d) \log P(X_{w'} \geq t_{w'}^d | \lambda_{w'})$$

L'équation 1 est un cas particulier de la formulation ci-dessus, obtenu en posant $\mathcal{F}(w) = \{w\}$, c'est-à-dire en considérant que les mots sont reliés seulement à eux-mêmes. L'application à un cadre multilingue consiste alors à utiliser la relation de

traduction, $\mathcal{F}(w) = TS(w)$, dans le calcul de $\mathcal{A}(w, w', d)$ ($TS(w)$ correspond à l'ensemble des traductions du mot w). Ceci conduit, pour les modèles LL et SPL, à :

$$RSV_{LL}(q,d) = -\sum_{w\in q} \frac{x_w^q}{l_q} \sum_{w'\in d} \mathcal{A}(w,w',d) \log(\frac{\lambda_{w'}}{\lambda_{w'} + t_{w'}^d}) \quad [3]$$

$$RSV_{SPL}(q,d) = -\sum_{w\in q} \frac{x_w^q}{l_q} \sum_{w'\in d} \mathcal{A}(w,w',d) \log(\frac{\lambda_{w'}^{\frac{t_{w'}^d}{t_{w'}^d+1}} - \lambda_{w'}}{1-\lambda_{w'}}) \quad [4]$$

Ces équations définissent deux nouveaux modèles de recherche d'information multilingue, que nous appellerons MI_{LL} et MI_{SPL}, MI signifiant *Mean Information*.

Une deuxième extension peut être obtenue en considérant que la variable aléatoire utilisée dans la famille des modèles d'information n'est pas associée à un mot seul w mais plutôt à un ensemble de mots, $\mathcal{F}(w)$, à savoir les mots reliés à w. Cela permet de définir une nouvelle fonction de recherche de la forme :

$$RSV(q,d) = -\sum_{w\in q} \frac{x_w^q}{l_q} \log P(X_{\mathcal{F}(w)} \geq t_{\mathcal{F}(w)}^d | \lambda_{\mathcal{F}(w)})$$

Comme précédemment, l'équation 1 est un cas particulier de cette fonction obtenu en posant $\mathcal{F}(w) = \{w\}$, et la version multilingue passe par la prise en compte de la relation de traduction : $\mathcal{F}(w) = TS(w)$. Il est néanmoins nécessaire ici de définir $t_{\mathcal{F}(w)}^d$ et $\lambda_{\mathcal{F}(w)}$. Nous posons simplement ici que $t_{\mathcal{F}(w)}^d$ équivaut à la somme des $t_{w'}^d$ des mots w' de $\mathcal{F}(w)$, ce qui correspond au fait que nous avons effectivement observé ce nombre d'occurrences (normalisées) de w dans d, au travers des mots qui lui sont reliés. La deuxième quantité ($\lambda_{\mathcal{F}(w)}$) est définie de façon similaire en considérant la fréquence documentaire normalisée de tous les mots de $\mathcal{F}(w)$ (cf. l'équation 2). Les formules suivantes formalisent ces deux définitions :

$$\begin{cases} t_{\mathcal{F}(w)}^d = & \sum_{w'\in\mathcal{F}(w)} t_{w'}^d \\ \lambda_{\mathcal{F}(w)} = & \frac{|\cup_{w'\in\mathcal{F}(w)} DS(w')|}{N} \end{cases}$$

Ce qui conduit à deux nouvelles formulations multilingues pour les modèles LL et SPL :

$$\text{RSV}_{LL}(q,d) = -\sum_{w\in q} \frac{x_w^q}{l_q} \log(\frac{\lambda_{\mathcal{F}(w)}}{t_{\mathcal{F}(w)}^d + \lambda_{\mathcal{F}(w)}}) \quad [5]$$

$$\text{RSV}_{SPL}(q,d) = -\sum_{w_f\in q} \frac{x_w^q}{l_q} \log(\frac{(\lambda_{\mathcal{F}(w)})^{\frac{t_{\mathcal{F}(w)}^d}{t_{\mathcal{F}(w)}^d+1}} - \lambda_{\mathcal{F}(w)}}{1-\lambda_{\mathcal{F}(w)}}) \quad [6]$$

L'extension ci-dessus est en partie analogue à l'opérateur SYN du système INQUERY, développé dans un cadre multilingue dans (Pirkola, 1998). En effet, cette formulation

peut aussi être dérivée en considérant que tous les mots reliés forment un seul et même mot. Nous avons montré ici comment la dériver d'autres considérations, à travers l'utilisation d'une seule varaiable aléatoire pour tous les mots reliés. La définition des paramètres associés (t_w^d and λ_w) découle alors du cadre général des modèles d'information. Pour cette raison, nous appelons les deux modèles de recherche d'information multilingue ci-dessus JV$_{LL}$ et JV$_{SPL}$, JV signifiant *Joint random Variable*.

Enfin, une troisième extension peut être obtenue en remplaçant tous les termes de la requête par l'ensemble de leurs traductions. Dans la mesure où la majorité des dictionnaires bilingues sont non pondérés, nous utilisons la formulation simple suivante, qui pourrait bien sûr être étendue par la prise en compte de poids de traduction :

$$RSV(q,d) = -\sum_{w \in q} \frac{x_w^q}{l_q} \sum_{w' \in d \cap TS(w)} \log P(X_{w'} \geq t_{w'}^d | \lambda_{w'})$$

Ceci conduit, pour les modèles LL et SPL, à :

$$RSV_{LL}(q,d) = -\sum_{w \in q} \frac{x_w^q}{l_q} \sum_{w' \in d \cap TS(w)} \log(\frac{\lambda_{w'}}{\lambda_{w'} + t_{w'}^d}) \quad [7]$$

$$RSV_{SPL}(q,d) = -\sum_{w \in q} \frac{x_w^q}{l_q} \sum_{w' \in d \cap TS(w)} \log(\frac{\lambda_{w'}^{\frac{t_{w'}^d}{t_{w'}^d+1}} - \lambda_{w'}}{1 - \lambda_{w'}}) \quad [8]$$

Nous appellerons les modèles ci-dessus QE$_{LL}$ et QE$_{SPL}$, QE signifiant *Query Expansion*.

En résumé, nous avons défini, au travers des développements précédents, trois versions multilingues des modèles LL and SPL models, dans le cadre général des modèles d'information :

1) MI$_{LL}$ and MI$_{SPL}$, corresponding to equations 3 and 4 ;

2) JV$_{LL}$ and JV$_{SPL}$, corresponding to equations 5 and 6 ;

3) QE$_{LL}$ and QE$_{SPL}$, corresponding to equations 7 and 8.

Avant d'évaluer ces modèles sur plusieurs collections de test, nous nous intéressons à la question d'une possible validation théorique. La section suivante est dédiée à cette question et s'appuie sur la théorie axiomatique de la recherche d'information.

2.2. *Validation théorique*

Un certain nombre de conditions portant sur les fonctions de recherche d'information ont été formulées dans (Fang *et al.,* 2004), Ce travail pionnier dans l'explicitation des conditions que les fonctions de recherche doivent satisfaire a été repris dans d'autres travaux, qui ont soit étendus soit reformulés ces conditions, comme par

exemple (Fang *et al.,* 2006, Cummins *et al.,* 2007, Clinchant *et al.,* 2010, Clinchant *et al.,* 2011, Zhai, 2011). Il y a maintenant de nombreux résultats empiriques qui montrent que les modèles de recherche d'information qui ne satisfont toutes les conditions standard ne sont pas optimaux. Comme montré dans (Clinchant *et al.,* 2010), les modèles LL et SPL que nous avons considéré satisfont toutes les conditions de la recherche d'information *ad hoc*. C'est également le cas pour les extensions que nous venons de voir. Cependant, le cadre multilingue apporte de nouveaux éléments par rapport à celui de la recherche d'information multilingue, et la question de l'existence de conditions propres à la recherche d'information multilingue se pose. Nous développons une telle condition ci-dessous.

Considérons d'une part une collection de documents français sur les rivières et les lacs, et d'autre part la requête anglaise *bank*. Dans ce contexte, les traductions possibles de *bank* en français sont *rive, berge, banc*[3]. Considérons maintenant que, dans un document d, les mots *berge* et *banc* apparaissent deux fois chacun et que, dans un autre document, d', le mot *rive* apparaît quatre fois. Supposons également que d et d' ont à peu près la même longueur, que *berge* et *banc* n'apparaissent que dans d et *rive* que dans d'. Toutes ces hypothèses peuvent être satisfaites dans une collection portant sur les cours d'eau et comportant des articles très formatés. Dans un tel contexte, il n'y a absolument aucune différence entre d et d' quand à leur pertinence par rapport à la requête, et une bonne stratégie multilingue devrait affecter le même score à ces deux documents. La condition suivante formalise cette intuition :

Condition 1 *Soient q une requête en langue source consistant en un seul mot w, d et d' deux documents en langue cible de même longueur. De plus, soit $\{w'_0, w'_1, \cdots, w'_k\}$ un ensemble de traductions de w, toutes ces traductions étant aussi probables les unes que les autres, tel que :*

$$\begin{cases} x^{d}_{w'_i} = 1, N_{w'_i} = 1, \ 1 \leq i \leq k \\ x^{d'}_{w'_0} = k, N_{w'_0} = 1 \end{cases}$$

Sous ces hypothèses, une bonne stratégie multilingue doit satisfaire :

$$RSV(q, d) = RSV(q, d')$$

Dans la mesure où la traduction de w est soit diluée sur plusieurs mots, dans d, soit concentrée sur un seul mot, dans d', nous appellerons la condition ci-dessus *condition DC*, DC signifiant Dilution/Concentration. Nous allons maintenant passer en revue les différents modèles multilingues que nous avons introduits au vu de cette condition. Nous nous concentrerons sur le modèle LL, le raisonnement et les résultats étant similaires pour le modèle SPL.

Comme toutes les traductions dans d ont le même nombre d'occurrences, elles ont aussi la même fréquence normalisée, que nous noterons $\tau : t^{d}_{w'_i} = \tau, \ 1 \leq i \leq k$. Nous

3. On peut penser à d'autres traductions possibles, mais cela ne change pas notre argumentation.

avons de plus : $t^{d'}_{w'_0} = k\tau$. Les hypothèses de la condition DC impliquent également que toutes les traductions ont le même paramètre λ : $\lambda_{w'_i} = \frac{1}{N},\ 0 \leq i \leq k$. Avec cela, nous avons :

– Pour l'extension QE :

$$RSV_{LL}(q,d) = k\log(\tau N + 1),\ RSV_{LL}(q,d') = \log(k\tau N + 1)$$

La fonction $RSV_{LL}(q,d') - RSV_{LL}(q,d)$ est strictement décroissante en τ, sa dérivée étant strictement négative, et égale à 0 quand $\tau = 0$, ce qui implique que :

$$RSV_{LL}(q,d') < RSV_{LL}(q,d)$$

La stratégie QE ne satisfait donc pas la condition DC.

– Pour l'extension MI :

$$RSV_{LL}(q,d) = \log(\tau N + 1),\ RSV_{LL}(q,d') = \log(k\tau N + 1)$$

Cette fois la fonction $RSV_{LL}(q,d') - RSV_{LL}(q,d)$ est croissante en τ pour $k \geq 1$, et égale à 0 quand $\tau = 0$, ce qui implique :

$$RSV_{LL}(q,d') > RSV_{LL}(q,d)$$

La stratégie MI ne satisfait donc pas la condition DC. On peut néanmoins remarquer que, dans cette stratégie, $\text{RSV}(q,d)$ est plus proche de $\text{RSV}(q,d')$ que dans la stratégie QE. En effet, en notant RSV_{QE} et RSV_{MI} les fonctions de recherche associées à ces stratégies, nous avons : $\text{RSV}_{QE}(q,d') = \text{RSV}_{MI}(q,d') = \text{RSV}(q,d')$. La fonction $\text{RSV}_{QE}(q,d) + \text{RSV}_{MI}(q,d) - 2\text{RSV}(q,d')$ est croissante en τ (sa dérivée est positive dès lors que $k\tau N > 1$, ce qui est le cas en pratique) et égale à 0 quand $\tau = 0$. Donc :

$$\text{RSV}_{QE}(q,d) - \text{RSV}_{QE}(q,d') > \text{RSV}_{MI}(q,d') - \text{RSV}_{MI}(q,d)$$

– Pour l'extension : $RSV_{JV}(q,d) = \log(k\tau N + 1) = RSV_{JV}(q,d')$. Cette extension satisfait donc la condition DC.

Le développement théorique ci-dessus montre que les extensions MI et QE ne satisfont pas la condition DC, la violation de cette condition étant moins forte pour l'extension MI. En revanche, l'extension JV satisfait la condition DC. Nous conjecturons donc que l'extension JV fournira de meilleurs résultats que l'extension MI, qui elle-même fournira de meilleurs résultats que l'extension QE. Comme nous allons le voir, les résultats expérimentaux sont en accord avec ces résultats théoriques.

3. Validation expérimentale

Nous utilisons dans nos expériences les collections anglaises avec les requêtes anglaises, françaises, allemandes et italiennes des campagnes CLEF bilingues, couvrant

les années 2000 à 2004[4]. Le tableau 2 fournit le le nombre de documents (N_d), le nombre de mots distincts (N_w), la longueur moyenne d'un document (DL_{avg}) de la collection anglaise de documents, ainsi que le nombre de requêtes, N_q, dans chaque tâche (toutes les requêtes sont disponibles dans les quatre langues mentionnées ci-dessus). Dans la mesure où les requêtes des années 2000 à 2002 portent sur la même collection, nous les avons rassemblées dans une même tâche. Les collections CLEF 2000-2002, CLEF 2003 et CLEF 2004 seront respectivement notées 02, 03 et 04 dans la suite. Dans toutes nos expériences, nous utilisons également des dictionnaires bilingues pour la traduction, comprenant respectivement 70 000k entrées pour le couple français-anglais, 58 000 entrées pour le couple allemand-anglais et 67 000 entrées pour le couple italien-anglais. La précision moyenne (*Mean Average Precision*), notée MAP, est utilisée pour évaluer la performance de chaque modèle. Enfin, le test T de Student par séries appariées (*paired t-test*), au niveau 0.05, est utilisé pour déterminer si les différences observées entre différents modèles sont significatives ou non.

Tableau 2. *Caractéristiques des différentes collections CLEF*

Collection	N_d	N_w	DL_{moy}	N_q
02	113,005	173,228	310.85	140
03	169,477	232,685	284.09	60
04	56,472	119,548	230.52	50

3.1. *Validation des extensions des modèles d'information*

Dans un premier temps, nous nous intéressons à la comparaison des différentes extensions (MI, JV et QE) proposées pour les modèles LL et SPL. Les modèles d'information comprennent un seul paramètre libre, le paramètre c, utilisée dans l'étape de normalisation des occurrences. Cette normalisation étant identique à celle utilisée dans les modèles DFR ((Amati *et al.,* 2002)), nous utilisons la valeur par défaut de ce paramètre fournie dans Terrier[5], à savoir $c = 1$. Les résultats obtenus pour les différentes collections et les trois couples de langue (c'est-à-dire français(fr)-anglais(ang), italien(it)-anglais(ang) et allemand(all)-anglais(ang)) sont résumés dans le tableau 3. Comme on peut le remarquer, et en accord avec le développement théorique de la section 2.2, l'extension JV fournit des résultats significativement meilleurs que les extensions MI et QE, pour les deux modèles LL et SPL. De même, l'extension MI fournit des résultats meilleurs que ceux obtenus avec l'extension QE. Dans la suite, nous n'utiliserons donc que l'extension JV pour les deux modèles LL et SPL de la famille des modèles d'information.

4. http ://www.clef-campaign.org
5. terrier.org

Tableau 3. *Comparaison des différentes extensions multilingues des modèles LL et SPL (la MAP sert ici de mesure d'évaluation). Le signe † indique, pour chaque modèle, que la différence avec le meilleur modèle (dont le score est donné en gras) est significative.*

Coll.		LL			SPL		
		JV	QE	MI	JV	QE	MI
02	fr-ang	**0.4174**	0.2042†	0.3748†	0.4008†	0.1937†	0.3702†
	it-ang	**0.3934**	0.2117†	0.3704†	0.3730†	0.1844†	0.3417†
	all-ang	**0.4102**	0.2124†	0.3750†	0.3901†	0.1990†	0.3574†
03	fr-ang	**0.4801**	0.2229†	0.4167†	0.4615†	0.2039†	0.4201†
	it-ang	**0.4339**	0.2133†	0.3817†	0.4210†	0.1991†	0.3746†
	all-ang	**0.4625**	0.2200†	0.3942†	0.4438†	0.2032†	0.3277†
04	fr-ang	**0.5204**	0.3085†	0.4171†	0.4317†	0.2317†	0.3460†
	it-ang	**0.4910**	0.2973†	0.4058†	0.4213†	0.2087†	0.3170†
	all-ang	**0.4921**	0.2969†	0.4062†	0.4222†	0.2166†	0.3276†

3.2. *Comparaison avec d'autres modèles multilingues*

Nous comparons maintenant les versions multilingues des modèles LL et SPL avec plusieurs modèles multilingues standard, à savoir : (a) le modèle vectoriel fondé sur les *tf* et l'*idf* de Robertson et Sparck-Jones ((Robertson *et al.,* 1988)), noté TF-IDF, (b) le modèle BM25 avec les valeurs par défaut des paramètres fournies par le système Terrier, (c) le modèle INQUERY avec les valeurs par défaut des paramètres fournies par le système Lemur[6] et (d) les modèles de langue Jelinek-Mercer et Dirichlet, avec les valeurs par défaut des paramètres fournies par le système Terrier ($\lambda = 0.15$ et $\mu = 2500$), notés LM-JM et LM-DIR. Pour les trois premiers modèles, nous utilisons la stratégie multilingue SYN qui consiste à considérer toutes les traductions d'un mot de la requête dans un document comme formant un seul et même mot. Il a en effet été montré que cette stratégie surpassait les stratégies alternatives dans plusieurs études ((Pirkola, 1998, Sperer *et al.,* 2000, Pirkola *et al.,* 2001, Ballesteros *et al.,* 2003)). Pour les modèles LM-JM et LM-DIR, deux autres stratégies ont été étudiées dans divers travaux (comme par exemple (Kraaij *et al.,* 2003)) : l'intégration des traductions dans le modèle de la requête (stratégie que nous noterons QT) ou dans le modèle du document (stratégie notée DT). Nous commençons par étudier ces différentes stratégies.

Le cas des modèles de langue

Les résultats de la comparaison des trois stratégies multilingues connues pour les modèles de langue (SYN, QT, DT) sont donnés dans le tableau 4, en terme de MAP et pour les trois couples de langue retenus. Comme nous pouvons le remarquer, la

6. www.lemurproject.org. Il n'existe pas à notre connaissance d'implantation de ce système dans Terrier.

stratégie SYN fournit des résultats significativement meilleurs que les autres stratégies sur tous les couples de langue et sur toutes les collections. Nous nous reposerons donc sur cette stratégie dans la suite.

Tableau 4. *Comparaison des différentes stratégies multilingues pour les modèles de langue (SYN, QT, DT). La MAP sert ici de mesure d'évaluation. Le signe* † *indique, pour chaque modèle, que la différence avec le meilleur modèle (dont les résultats sont en gras) est significative. Pour des raisons de lisibilité, quand cette différence n'est pas significative, les résultats sont en italique.*

Coll.		DT		QT		SYN	
		JM	DIR	JM	DIR	JM	DIR
02	fr-ang	0.3711†	0.3924†	0.3641†	0.3491†	0.3930†	**0.4102**
	it-ang	0.3497†	0.3660†	0.3207†	0.3143†	0.3720†	**0.3878**
	all-ang	0.3728†	0.3797†	0.3490†	0.3504†	*0.3925*	**0.3983**
03	fr-ang	0.4419†	0.4038†	0.3981†	0.3781†	**0.4716**	0.4242†
	it-ang	0.4162†	0.4211†	0.3745†	0.3801†	**0.4355**	0.3857†
	all-ang	0.4271†	0.3713†	0.3813†	0.3336†	**0.4554**	0.4098†
04	fr-ang	0.4217†	0.4222†	0.3861†	0.4186†	**0.4513**	*0.4417*
	it-ang	0.3907†	0.3824†	0.3778†	0.3812†	**0.4221**	*0.4201*
	all-ang	0.3992†	0.3874†	0.3810†	0.3796†	**0.4331**	*0.4310*

Il est aussi intéressant de remarquer que la stratégie DT fournit des résultats qui sont, de façon systématique, meilleurs que ceux obtenus par la stratégie QT, qui est la plus mauvaise des stratégies ici. La stratégie QT est en fait la seule qui ne vérifie pas la condition DC introduite dans la section 2.2. Il est en effet direct de vérifier que la stratégie SYN satisfait la condition DC, que ce soit pour le modèle LM-JM ou pour le modèle LM-DIR, car, dans cette stratégie, les différents traductions dans d sont regroupées sous une même forme ayant k occurences ; les documents d et d' deviennent dès lors identiques du point de vue de la traduction. Pour la stratégie DT, avec les hypothèses de la condition DC, nous obtenons :

$$\text{RSV}_{DT}(q, d') = \log(P(w|w'_0)P(w'_0|\mathcal{M}_{d'}))$$

ainsi que :

$$\text{RSV}_{DT}(q, d) = \log(\sum_{i=1}^{k} P(w|w'_i)P(w'_i|\mathcal{M}_d)) \quad [9]$$

L'utilisation du lissage de Jelinek-Mercer avec un paramètre de lissage λ conduit à :

$$\begin{aligned} P(w'_0|\mathcal{M}_{d'}) &= k((1-\lambda)\frac{1}{l_{d'}} + \lambda\frac{1}{L}) \\ &= kPw'_i|\mathcal{M}_d),\ (1 \leq i \leq k) \end{aligned} \quad [10]$$

De plus, toujours sous les hypothèses de la condition DC nous avons :

$$P(w|w'_0) = P(w|w'_i),\ (1 \leq i \leq k) \quad [11]$$

En combinant les équations 11, 10 et 9, nous obtenons : $\text{RSV}_{DT}(q, d') = \text{RSV}_{DT}(q, d)$, ce qui montre que le modèle LM-JM avec la stratégie DT satisfait la condition DC. Le même développement peut être mené sur le modèle LM-DIR, qui satisfait lui aussi la condition DC. La situation est toutefois différente pour la stratégie QT, pour laquelle nous avons :

$$\text{RSV}_{QT}(q, d') = P(w'_0|w) \log(P(w'_0|\mathcal{M}_{d'}))$$

et :

$$\begin{aligned}
&\text{RSV}_{QT}(q, d) \\
&= \sum_{i=1}^{k} (P(w'_i|w) P_{ML}(w|\mathcal{M}_q) \log(P(w'_i|\mathcal{M}_d))) \\
&= kP(w'_i|w) \log(P(w'_i|\mathcal{M}_d)),\ (1 \leq i \leq k)
\end{aligned}$$

Avec le lissage de Jelinek-Mercer, ces quantités prennent la forme :

$$\text{RSV}_{QT}(q, d') - \text{RSV}_{QT}(q, d) = P(w'_i|w)(\log k - (k-1) \log(P(w'_i|\mathcal{M}_d))) \quad [12]$$

une quantité plus grande que $P(w_e^i|w_f) \log k$, et donc plus grande que 0 dès que $k \geq 1$. Le modèle LM-JM avec la stratégie QT ne satisfait donc pas la condition DC. Ici encore le même développement peut être mené sur le modèle LM-DIR, avec la même conclusion.

Comparaison générale

Enfin, le tableau 5 fournit les résultats obtenus avec les différents modèles multilingues que nous avons retenus, sur tous les couples de langue et toutes les collections (pour les modèles d'information, nous ne présentons que le modèle LL, qui fournit les meilleurs résultats ici). La ligne MON correspond à la version monolingue de ces différents modèles. Comme on peut le constater, le modèle LL_{JV} soit fournit les meilleurs résultats (dans 9 cas sur 12), soit la différence avec le meilleur modèle n'est pas significative. De plus, quand ce modèle obtient le meilleur score, la différence avec les autres modèles est significative dans la plupart des cas. En effet, seul le modèle LM-DIR est comparable au modèle LL_{JV} sur la collection 2000-2002 collection, seul le modèle LM-JL est comparable avec LL_{JV} sur la collection 2003, et tous les modèles sont en dessous de LL_{JV} sur la collection 2004, la différence étant significative.

4. Conclusion

Nous avons présenté ici plusieurs extensions multilingues des modèles d'information à travers (a) une généralisation de la notion d'information utilisée dans ces modèles, (b) une généralisation des variables aléatoires utilisées et (c) une expansion de la requête utilisant l'ensemble

Tableau 5. *Comparaison des différents modèles multilingues sur la MAP, pour tous les couples de langue et toutes les collections. Le signe † indique, pour chaque modèle, que la différence avec le meilleur modèle (dont les résultats sont en gras) est significative. Pour des raisons de lisibilité, quand cette différence n'est pas significative, les résultats sont en italique.*

Data	Model	TF-IDF	BM25	LM-JM	LM-DIR	INQUERY	LL_{JV}
02	MON	0.4475†	0.4744†	0.4621†	0.4783†	0.4227†	**0.4866**
	fr-ang	0.3641†	0.3891†	0.3990†	*0.4102*	0.3527†	**0.4174**
	it-ang	0.3325†	0.3578†	0.3720†	*0.3878*	0.3216†	**0.3934**
	all-ang	0.3502†	0.3674†	*0.3925*	*0.3983*	0.3419†	**0.4102**
03	MON	0.4763†	**0.5031**	0.4919†	0.4751†	0.4369†	*0.5030*
	fr-ang	0.4155†	0.4405†	*0.4716*	0.4242†	0.4076†	**0.4801**
	it-ang	0.3764†	0.4000†	**0.4355**	0.3857†	0.3732†	*0.4339*
	all-ang	0.3966†	0.4198†	*0.4554*	0.4098†	0.3842†	**0.4625**
04	MON	0.5187†	0.5228†	0.5110†	**0.5386**	0.4264†	*0.5381*
	fr-ang	0.4225†	0.4197†	0.4513†	0.4417†	0.3763†	**0.5204**
	it-ang	0.3917†	0.3834†	0.4221†	0.4201†	0.3425†	**0.4910**
	all-ang	0.4008†	0.3947†	0.4331†	0.4310†	0.3534†	**0.4921**

des traductions de chaque mot. L'approche fondée sur la généralisation des variables aléatoires est analogue à la stratégie SYN utilisée dans plusieurs travaux antérieurs. Le bon comportement de cette stratégie, déjà noté dans ces travaux antérieurs, est confirmé ici à la fois d'un point de vue théorique et expérimental.

Nous avons de plus introduit une nouvelle condition pour la recherche d'information multilingue, étendant ainsi l'analyse axiomatique de la recherche d'information. Cette nouvelle condition, que nous appelons condition de Dilution/Concentration (DC), nous a permis d'évaluer d'un point de vue purement théorique les extensions multilingues des modèles d'information que nous avons introduites. Les résultats obtenus par cette évaluation théorique ont été confirmés par les résultats expérimentaux. Nous avons également utilisé la condition DC pour évaluer, toujours d'un point de vue théorique, les stratégies multilingues proposées pour les modèles de langue. Ici encore, cette évaluation théorique a été confirmée par nos évaluations expérimentales.

Enfin, nous avons montré que l'extension multilingue du modèle log-logistique fondée sur la généralisation des variables aléatoires (analogue à la stratégie SYN) fournissait les meilleurs résultats de recherche d'information multilingue sur trois collections et trois couples de langue. EN effet, ce modèle n'est jamais "dominé" de façon significative par un autre modèle, mais en revanche fournit en général des résultats significativement meilleurs que ceux obtenus avec d'autres modèles. Sa forme simple, donnée par l'équation 5, en fait de plus un modèle attractif d'un point de vue opérationnel.

Dans le futur, nous comptons poursuivre nos expériences en optimisant les paramètres des différents modèles. Nous nous sommes reposés dans cette étude sur les valeurs par défaut de ces paramètres (fournies par Lemur et Terrier) car ces valeurs sont couramment utilisées lorsque de nouvelles collections et requêtes doivent être traitées. C'est également ces valeurs qui sont utilisées par beaucoup de participants aux campagnes d'évaluation de recherche d'information

multilingue qui désirent utiliser les systèmes de recherche comme une boîte noire et se concentrer sur le développement d'outils de pré- ou post-traitement.

5. Bibliographie

Amati G., Van Rijsbergen C. J., « Probabilistic models of information retrieval based on measuring the divergence from randomness », *ACM Trans. Inf. Syst.*, vol. 20, p. 357-389, October, 2002.

Ballesteros L., Sanderson M., « Addressing the lack of direct translation resources for cross-language retrieval », *Proceedings of the twelfth international conference on Information and knowledge management*, CIKM '03, New Orleans, LA, USA, p. 147-152, 2003.

Braschler M., « Combination Approaches for Multilingual Text Retrieval », *Inf. Retr.*, vol. 7, p. 183-204, January, 2004.

Clinchant S., Gaussier E., « Information-based models for ad hoc IR », *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, ACM, New York, NY, USA, p. 234-241, 2010.

Clinchant S., Gaussier É., « Is Document Frequency Important for PRF ? », *ICTIR*, p. 89-100, 2011.

Cummins R., O'Riordan C., « An axiomatic comparison of learned term-weighting schemes in information retrieval : clarifications and extensions », *Artif. Intell. Rev.*, vol. 28, p. 51-68, June, 2007.

Fang H., Tao T., Zhai C., « A Formal Study of Information Retrieval Heuristics », *SIGIR '04 : Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004.

Fang H., Zhai C., « Semantic term matching in axiomatic approaches to information retrieval », *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, p. 115-122, 2006.

Federico M., Bertoldi N., « Statistical cross-language information retrieval using n-best query translations », *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, Tampere, Finland, p. 167-174, 2002.

Kraaij W., Nie J.-Y., Simard M., « Embedding web-based statistical translation models in cross-language information retrieval », *Computational Linguistic*, vol. 29, p. 381-419, September, 2003.

Nie J.-Y., *Cross-Language Information Retrieval*, Morgan & Claypool, New York, NY, USA, 2010.

Pirkola A., « The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval », *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, ACM, New York, NY, USA, p. 55-63, 1998.

Pirkola A., Hedlund T., Keskustalo H., Järvelin K., « Dictionary-Based Cross-Language Information Retrieval : Problems, Methods, and Research Findings », *Inf. Retr.*, vol. 4, p. 209-230, September, 2001.

Robertson S. E., Sparck Jones K., *Relevance weighting of search terms*, p. 143-160, 1988.

Sperer R., Oard D. W., « Structured translation for cross-language information retrieval », *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, Athens, Greece, p. 120-127, 2000.

Zhai C., « Axiomatic Analysis and Optimization of Information Retrieval Models », *ICTIR*, 2011.