Publié dans les
Actes 8e Conférence en Recherche d'Information et Applications CORIA'11,
271-286, 2011
qui doivent être utilisés pour toute référence à ce travail

Recherche d'information dans un corpus bruité (OCR)

Nada Naji, Jacques Savoy, Ljiljana Dolamic

Institut d'informatique
Université de Neuchâtel, rue Emile Argand 11, 2000 Neuchâtel (Suisse)
{Nada.Naji, Jacques.Savoy, Ljiljana.Dolamic}@unine.ch

RÉSUMÉ. *Cet article désire mesurer la perte de performance lors de la recherche d'information dans une collection de documents scannés. Disposant d'un corpus sans erreur et de deux versions renfermant 5 % et 20 % d'erreurs en reconnaissance, nous avons évalué six modèles de recherche d'information basés sur trois représentations des documents (sac de mots, n-grammes, ou trunc-n) et trois enracineurs. Basé sur l'inverse du rang du premier document pertinent dépisté, nous démontrons que la perte de performance se situe aux environs de - 17 % avec un taux d'erreur en reconnaissance de 5 % et s'élève à – 46 % si ce taux grimpe à 20 %. La représentation par 4-grammes semble apporter une meilleure qualité de réponse avec un corpus bruité. Concernant l'emploi ou non d'un enracineur léger ou la pseudo-rétroaction positive, aucune conclusion définitive ne peut être tirée.*

ABSTRACT. *This paper evaluates the retrieval effectiveness degradation when facing with noisy text corpus. With the use of a test-collection having the clean text, another version with around 5% error rate in recognition and a third with 20% error rate, we have evaluated six IR models based on three text representations (bag-of-words, n-grams, trunc-n) as well as three stemmers. Using the mean reciprocal rank as performance measure, we show that the average retrieval effectiveness degradation is around -17% when dealing with an error rate of 5%. This average decrease is around -46% when facing with an error rate of 20%. The representation by 4-grams tends to offer the best retrieval when searching with noisy text. Finally, we are not able to obtain clear conclusion about the impact of different stemming strategies or the use of blind-query expansion.*

MOTS-CLÉS : *Recherche d'information dans des documents bruités (OCR), évaluation, TREC.*

KEY WORDS: *Information retrieval with noisy text (OCR), Evaluation, TREC.*

1. Introduction

Depuis une dizaine d'années les diverses sources de notre héritage culturel font l'objet de nombreuses études afin de mieux les préserver, d'y accéder plus aisément ou de pouvoir procéder à des traitements informatiques. Les projets les plus

importants possèdent un ancrage essentiellement national (e.g., la bibliothèque numérique *Gallica*) ou international (*The European Library*, ou *Europeana*[1]). Si la conservation constitue souvent le point de départ, Internet a démontré l'intérêt de promouvoir un accès à distance pour des sources dont la manipulation s'avère délicate. Pour les informaticiens, les défis sont multiples comme la conception d'interface intuitive, conviviale et performante, la gestion de supports multiples (textes, manuscrits, peintures, cartes, musiques, bandes son, vidéo, etc.) ou la prise en compte des divers aspects multilingues.

Dans ce cadre, notre groupe participe au projet HisDoc[2] visant d'abord à digitaliser des sources manuscrites médiévales. Ensuite, nous désirons procéder à une reconnaissance automatique d'une sélection assez large de ces manuscrits écrits, essentiellement en langue allemande. Enfin, le projet HisDoc disposera d'un moteur de recherche pour permettre un accès aisé aux sources retenues. Dans ce dernier aspect, notre premier objectif consiste à évaluer les diverses stratégies d'indexation et de recherche d'information en présence de documents bruités (reconnaissance optique imparfaite des caractères (OCR) ou orthographe non normée).

Dans la suite de cette communication, nous désirons présenter quelques études reliées à l'accès à des documents issus d'une reconnaissance imparfaite des caractères (section 2). La troisième section expose les grandes lignes du corpus utilisé dans nos expériences. La quatrième section décrit les stratégies d'indexation et de recherche retenues. La cinquième section présente notre méthodologie d'évaluation et l'appliquera à diverses stratégies d'indexation et de dépistage.

2. Défis sous-jacents et état de l'art

Peu d'études ont été entreprises sur une échelle assez large afin de connaître la perte d'efficacité provenant d'une reconnaissance défaillante des caractères [MIC 00], [CAL 02]. Les campagnes TREC-3 à TREC-5 (*confusion track*) constituent un point de départ intéressant [KAN 00], [VOO 05]. Les expériences menées durant la campagne TREC-4 ne reflètent pas très précisément nos objectifs puisque les documents ont été corrompus de manière aléatoire au NIST (sans refléter une source précise d'erreur). Lors de la campagne TREC-5, trois versions différentes d'un corpus écrit en langue anglaise (55 600 documents) ont été mises à disposition des participants. La première version, ne renfermant pas d'erreur, va servir de base de référence. La deuxième version des documents a été obtenue après scannage et reconnaissance optique des caractères. Dans ce cas, le taux d'erreur se situe à environ 5 %. La dernière version est similaire à la précédente mais avec un taux d'erreur de l'ordre de 20 %.

---

[1] Voir les sites http://gallica.bnf.fr/, http://search.TheEuropeanLibrary.org/, ou http://www.europeana.eu/.

[2] Voir le site http://hisdoc.unine.ch/.

Afin de mesurer la qualité de la recherche d'information en présence de documents bruités, la campagne d'évaluation TREC a eu recours au MRR (*mean reciprocal rank*), une mesure axée vers la précision et basée uniquement sur l'inverse du rang du premier document pertinent retrouvé [ABD 07]. Une telle mesure reflète bien les préoccupations d'un internaute désirant trouver une ou un nombre limité de bonnes réponses à sa requête.

A l'aide de cette mesure, le meilleur système de TREC-5 [BAL 97] possédait une valeur MRR de 0,7353 sur les documents sans erreur, une valeur MRR de 0,5737 (différence relative de -22 %) avec un taux d'erreur de 5 % ou une valeur MRR de 0,4978 (-32 %) en présence d'un taux d'erreur de 20 %. Des niveaux de dégradation similaires ont été obtenus par d'autres participants [VOO 05]. On peut toutefois relativiser ces premiers résultats. Par exemple, Tagva *et al*. [TAG 94] ont démontré qu'en présence d'images de haute définition et d'un système OCR de haute qualité, le taux d'erreur peut se limiter aux environs de 2 %.

Comment expliquer les effets de données bruitées sur la qualité de réponse d'un moteur de recherche ? Tout d'abord, on doit reconnaître que toute communication en langue naturelle se fonde sur une certaine redondance. Ainsi, il est rare qu'un même mot soit mal orthographié plusieurs fois. Pour des taux d'erreur relativement faibles, cette redondance permet au système de recherche de retrouver les éléments souhaités. Mais si l'erreur touche un nom propre apparaissant une ou deux fois, l'effet sera plus important. Si l'on analyse les termes d'indexation provenant d'un corpus OCR, on constate qu'ils tendent à former un sur-ensemble des termes extraits de la même collection [TAG 04]. L'OCR va produire plus de termes ayant une fréquence d'occurrence unitaire. Selon Tagva *et al*. [TAG 94] cette proportion peut atteindre 70 %, soit une valeur plus élevée que le niveau théorique de 50 % prévu par la loi de Zipf ou que ceux observés dans des études empiriques (avec des valeurs entre 38 % et 44 %, selon les hommes politiques et leurs discours [SAV 10]).

La présence d'un nombre restreint de termes mal orthographiés peut modifier les poids lors de l'indexation et de la recherche, mais de telles différences ne devraient pas générer une baisse sensible de la performance globale. En revanche, si le taux d'erreur grandit, la qualité de la réponse s'en ressent et diminue nettement. En effet, en présence de documents bruités, la fréquence documentaire tend à être plus réduite, générant ainsi des valeurs *idf* plus élevées. Dès lors, la normalisation par la fonction cosinus n'apporte pas l'effet souhaité et une autre forme de normalisation devrait être appliquée. Pour Mittendorf & Schäuble [MIS 00], le principal problème ne réside pas dans les erreurs de reconnaissance en elles-mêmes. Toutefois, leurs présences dans un document relativement bref peuvent avoir un effet dévastateur.

Si l'on analyse d'autres composantes d'un système de recherche en présence de données bruitées, aucune conclusion claire et définitive ne semble être possible. Par exemple, Taghva *et al*. [TAG 94] suggèrent d'ignorer tout enracineur (*stemmer*) ou, pour le moins, d'appliquer uniquement un enracineur léger (éliminant uniquement la marque du pluriel pour la langue anglaise). En revanche le meilleur système à

TREC-5 [BAL 97] utilisait un enracineur agressif (Porter [POR 80] dans le cas présent). De plus, ce système de recherche ignorait la plupart des mots possédant moins de trois lettres (e.g., “in,” “the”, “by”). Sachant que la plupart des erreurs de reconnaissance touchent les mots ayant deux ou trois lettres, l'élimination systématique de ces formes brèves mérite d'être envisagée. Finalement, l'expansion automatique des requêtes (*blind-query expansion*) tend à apporter une amélioration moindre qu'en présence de documents correctement orthographiés [TAG 96].

D'autres auteurs suggèrent de procéder à une correction automatique des données bruitées avant l'indexation, en se basant sur une modélisation de la source d'erreur (e.g., en recourant à la distance d'édition). Naturellement, plusieurs méthodes de correction ont été proposées dans cette perspective [PIL 06], en s'appuyant parfois sur le Web ou Wikipedia pour identifier des orthographes potentiellement correctes [CUC 04]. Cependant, si l'on travaille sur des langues moins populaires ou anciennes, le recours au Web n'a qu'une valeur limitée.

Corriger automatiquement tout le corpus s'avère habituellement impossible. En revanche, un processus de correction automatique peut se limiter aux termes de la recherche, une stratégie retenue par les principaux moteurs commerciaux. Comme autre exemple, Ballerini *et al*. [BAL 97] suggèrent de corriger les erreurs seulement pour les m (e.g., m = 2 000) documents les mieux classés plutôt que de considérer tous les articles dépistés. Dans un tel cas de figure, la valeur *idf* n'est plus identique à celle obtenue sur la base des index du système. Comme mesure palliative, on peut procéder à une estimation des valeurs *idf* sur la base d'un corpus comparable voire de recourir à une mesure *idf* révisée [MIS 00].

Finalement, l'indexation peut s'appuyer sur des séquences consécutives de n-caractères [MCN 04] plutôt que sur des mots. Ballerini *et al*. [BAL 97] suggèrent de recourir à des séquences superposées de trigrammes ou de quadri-grammes. En présence d'un taux d'erreur d'environ 10 %, une indexation par des n-grammes tend à apporter une meilleure qualité (+ 10 % en précision moyenne) qu'une approche basée sur les mots [HAR 97]. Dans cette dernière étude, l'indexation proposée était basée sur des bigrammes et trigrammes, ou selon une combinaison de bigrammes, trigrammes, 4-grammes, et même 5-grammes.

En résumé, ces diverses études démontrent que la qualité de la recherche d'information tend à diminuer en fonction du taux d'erreur en reconnaissance. Basée sur les expériences TREC, cette diminution relative atteint en moyenne les - 38 % pour un taux d'erreur situé à 5 % et de - 52 % pour un taux d'erreur d'environ 20 %. On notera toutefois que ces études s'appuient essentiellement sur des corpus écrits en langue anglaise, utilisant des documents soit dactylographiés ou écrits durant les deux derniers siècles. La situation peut devenir plus complexe face à d'autres langues ayant une morphologie plus complexe ou lorsque les documents manuscrits proviennent d'une époque où l'orthographe n'était pas normalisée [DEC 09].

3. Collection utilisée dans nos expériences

La collection utilisée pour nos évaluations a été créée lors des campagnes d'évaluation TREC-4 et TREC-5. Elle comprend un ensemble de 55 630 documents dont 125 entrées correspondent à des articles sans texte. Ces articles ont été extraits du *Federal Register* (durant l'année 1994), une publication quotidienne contenant les règlements ou les propositions de règlements, ainsi que diverses informations provenant des agences et organisations fédérales américaines[3]. En moyenne chaque document comprend 414,05 termes d'indexation (ou, en moyenne, 202,5 termes distincts).

```
<DOC>
<DOCNO> FR940317-1-00199
<TEXT>
In withdrawing the riskless principal mark-up disclosure
proposal in the 1978 Release, the Commission stated that it
would ''maintain close scrutiny to prevent excessive mark-ups
and take enforcement action where appropriate.''
Since 1982, the Commission and NASD have undertaken a number
of enforcement actions against broker-dealers involving …

<DOC>
<DOCNO> FR940317-1-00199
<TEXT>
ln withdrawlng the risyless principal mary-up disclosure
proposal in the 191W helease1 the Commission stated that it
would 44maintain close scrutiny to prevent excessive mary-ups
and taye enforcement action where appropriate.:: 20
Since 19W21 the Commission and NASL have undertayen a number
of enforcement actions against broyer-dealers involving …

<DOC>
<DOCNO> FR940317-1-00199
<TEXT>
fa -thtlrawing the WfUefqs priucipA mary-up dRclosure proposA
in the 191@ M,lease, the ComMssioa stated that it would
amUntdn close scrutAy to preveat excessive m=y-upqe at nd
tttes eaforcemebt actioa where approphate.. 2e 0
Since 19S2, tlte CommLsion and htASO have .ndertayea a a0mber
of eaforc,ement actioaf agUnst broyer-de,Uers involUng …
```

Figure 1 *: Exemple du même passage en version originale, puis avec 5 % et 20 % d'erreur de reconnaissance*

La figure 1 présente un exemple de document extrait de ce corpus. Chaque document débute par la balise <DOC>, avec comme deuxième balise l'identificateur

[3] Le site Internet www.gpoaccess.gov/fr/ permet d'accéder à son contenu.

unique de l'article (<DOCNO>), puis le texte proprement-dit (étiquette <TEXT>). Comme le démontre ces courts extraits, la présence d'un taux d'erreur de 5 % en reconnaissance permet tout de même une lecture de l'article. En revanche, avec un taux s'élevant à 20 %, la lecture continue s'avère difficile et l'on procède plus à un déchiffrement qu'à une lecture. On peut dès lors s'imaginer les problèmes de représentation et de dépistage de l'information sur un tel corpus.

Avec ces documents, nous disposons de 50 requêtes numérotées de 1 à 50 (corpus TREC-5). Toutefois, aucune bonne réponse pour la requête nº 29 n'existe dans le corpus; cette interrogation a donc été éliminée dans les évaluations. Pour les autres, la collection dispose d'une seule et unique bonne réponse par requête. Ce type de recherche correspond à la recherche d'un objet connu (*known-item search*). L'expérience simule le besoin d'un usager qui se souvient partiellement d'un texte qu'il a déjà vu et qu'il désire retrouver. Chaque requête contient un seul champ utile (suivant la balise <DESC>, voir exemples donnés en figure 2).

```
<NUM>  CF2
<DESC>  Mutual fund risk disclosure for futures related issues.

<NUM>  CF5
<DESC>  National Science Programs in radio astronomy.

<NUM>  CF30
<DESC>  characterization and symptoms of the milk-alkali
syndrome.
```

***Figure 2** : Exemple de requêtes de notre corpus*

Les thèmes des demandes couvrent des domaines variés comme la science (nº 4 "Use of optical character recognition for creation of electronic databases"), l'économie et la finance (nº 7 "Excessive mark up of zero coupon treasury bonds"), la santé (nº 47 "What areas of the human body are the main targets for lead?"), la politique (nº 9 "Gold watches accepted as gifts to the US government."), la science et la technologie (nº 10 "Patent applications for nonlinear neural network oscillators") et les faits divers (nº 13 "I am looking for theft data on the Chevrolet Corsica."). Notons que certaines requêtes se limitent à un seul mot (nº 35 "bugle") ou à un groupe de deux ou trois mots (nº 34 "saucer bounce", nº 39 "jellyfish abundance").

4. Les stratégies d'indexation et modèles de recherche d'information

Dans cet article, nous désirons obtenir une vision assez large de la performance obtenue par diverses combinaisons de représentations des documents et de modèles de recherche de l'information [MAN 08]. Comme première représentation des articles et des requêtes, nous proposons de recourir au paradigme du sac de mots. Après une segmentation en mots, le système élimine les formes très fréquentes et

peu ou pas porteuses de sens (mots-outils). Dans le cadre de cette étude, cette liste comprend 571 entrées (liste du système SMART). Ensuite nous pouvons procéder à l'élimination automatique de certaines séquences terminales. Dans ce but, nous pouvons appliquer un enracineur léger [HAR 91] supprimant la marque du pluriel en langue anglaise (correspondant à l'élimination de la consonne finale '-s', règles indiquées dans la table 1). Comme variante, l'algorithme plus agressif proposé par Porter [POR 80] supprime également certains suffixes dérivationnels. Cependant, ce choix entraîne le risque de réduire sous la même forme des mots de sens assez différents (e.g., "organisation" devient "organ") et d'ignorer certains regroupements pertinents (e.g., "merger" et "merging" ne sont pas réunis sous la même entrée).

```
Si la finale est '-ies' mais pas '-eies' ou '-aies'
        alors remplacez  '-ies' par '-y', fin;
Si la finale est '-es' mais pas '-aes', '-ees' ou '-oes'
        alors remplacez  '-es' par '-e', fin;
Si la finale est '-s' mais pas '-us' ou '-ss'  alors éliminez  '-s';
fin.
```

Table 1 *: Les trois règles de l'enracineur léger suggéré par Harman [HAR 91]*

Comme représentation alternative, nous pouvons également recourir à des séquences consécutives de n lettres. Si l'on fixe $n = 4$, le mot "computing" générera les termes d'indexation "comp", "ompu", …, "ting". Avec cette approche, nous ignorons tout processus d'élimination de certaines séquences finales, laissant au schéma de pondération le soin d'attribuer des valeurs très faibles aux séquences correspondant à des suffixes fréquents de la langue (e.g., "ting", or "ably"). Nous avons fixé la valeur $n = 4$ car elle nous permet d'obtenir les meilleures performances.

Comme autre choix, nous suggérons de tenir compte uniquement des n premières lettres de chaque mot (troncature à n). Si l'on fixe $n = 4$, le mot "computing" générera le terme d'indexation "comp". Face à un taux d'erreur en reconnaissance élevé, la chance pour qu'un mot long ne soit pas altéré demeure faible. Ainsi, avec un taux d'erreur de 20 %, la probabilité qu'une séquence de trois lettres soit préservée s'élève à $(1 - 0{,}2)^3 = 0{,}512$ et seulement à $(1 - 0{,}2)^4 = 0{,}41$ si l'on considère une séquence de quatre lettres. Ce calcul semble indiquer qu'une valeur de n faible (trois ou moins) serait adéquate. Cependant, une valeur trop faible (e.g., $n = 2$) risque d'induire un nombre trop important d'appariements erronés. En effet, le dépistage s'effectuera entre les n premières lettres des termes de la requête et ceux des documents, laissant la place à un nombre élevé d'appariements non justifiés sémantiquement. De plus, on peut tenir compte d'une certaine redondance dans les documents permettant de penser qu'un mot significatif sera présent plusieurs fois, réduisant ainsi la probabilité de l'absence de tout appariement.

Les termes extraits des documents et des requêtes étant définis, nous pouvons les pondérer selon la formulation classique $tf \cdot idf$. Dans ce cas, nous tenons compte de

la fréquence d'occurrence (ou fréquence lexicale notée tf_{ij} pour le j^e terme dans le i^e document) et de la fréquence documentaire d'un terme (df_j, ou plus précisément de l'$idf_j = \log(n/df_j)$ avec n indiquant le nombre de documents inclus dans le corpus). Pour être précis, notre pondération inclut également une normalisation selon le cosinus [SAV 08].

D'autres variantes ont été proposées dans le cadre du modèle vectoriel, en particulier en considérant que l'apparition d'un nouveau terme doit être perçue comme un événement plutôt rare. Dans ce cas, on doit attacher plus d'importance à la première occurrence comparée aux suivantes. Dès lors la composante *tf* peut s'évaluer comme $\ln(tf) + 1$ ou selon $\ln(\ln(tf)+1)+1$. D'autre part, la présence d'un terme dans un article bref doit être représenté par une pondération plus forte que sa présence dans un long document. Afin de tenir compte des différences de longueur, Buckley *et al*. [BUC 96] suggèrent de recourir à la pondération *Lnu-ltu* dans laquelle *Lnu* correspond à la pondération des termes dans les documents (voir équation 1) et *ltu* à celle des termes de la requête (voir équation 2). Cette formulation a été choisie par le meilleur système à TREC-5 [BAL 97].

$$w_{ij} = \frac{\left(\left(\ln(tf_{ij})+1\right) \Big/ \left(\ln(\text{mean tf})+1\right)\right)}{(1-slope)\cdot pivot + slope \cdot nt_i} \quad (1)$$

$$w_{qj} = \frac{\left(\ln(tf_{qj})+1\right)\cdot idf_j}{(1-slope)\cdot pivot + slope \cdot nt_i} \quad (2)$$

Dans les formules ci-dessus, nt_i indique le nombre de termes d'indexation distincts dans le document d_i tandis que *pivot* et *slope* sont deux constantes utilisées pour normaliser les poids en fonction de la longueur moyenne des documents. Ce modèle vectoriel s'avère particulièrement intéressant lorsque l'on cherche à obtenir une précision élévée pour les premiers documents dépistés [DOL 10].

Ces deux premiers modèles vectoriels seront complétés par des approches probabilistes. Dans ce cadre, nous avons considéré le modèle Okapi [ROB 00] utilisant la formulation suivante :

$$w_{ij} = [(k_1+1)\cdot tf_{ij}] / (K + tf_{ij}) \quad \text{avec } K = k_1 \cdot [(1-b) + ((b \cdot l_i) / mean\ dl)] \quad (3)$$

formule dans laquelle l_i est la longueur du i^e article (nombre de termes d'indexation), et b, k_1, *mean dl* des constantes fixées à $b = 0{,}55$, $k_1 = 1{,}2$ et *mean dl* = 414.

Comme deuxième modèle probabiliste, nous avons implémenté le modèle PL2, un des membres de la famille *Divergence from Randomness* (DFR) [AMA 02]. Dans ce cas, la pondération w_{ij} combine deux mesures d'information, à savoir :

$$\begin{aligned} &w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = Inf^1_{ij} \cdot (1 - Prob^2_{ij}) \quad \text{et} \\ &Prob^2_{ij} = tfn_{ij} / (tfn_{ij} + 1) \quad \text{avec } tfn_{ij} = tf_{ij} \cdot \ln[1+((c \cdot mean\ dl) / l_i)] \\ &Inf^1_{ij} = -\log_2[(e^{-\lambda_j} \cdot \lambda_j^{tfn_{ij}})/tfn_{ij}!] \quad \text{avec } \lambda_j = tc_j / n \end{aligned} \quad (4)$$

dans laquelle tc_j représente le nombre d'occurrences du j^e terme dans la collection, n le nombre d'articles dans le corpus et c une constante fixée à 1,5.

Comme troisième modèle probabiliste, nous avons retenu le modèle I(n_e)B2 également issu de la famille DFR se basant sur la formulation suivante :

$$\text{Prob}^2_{ij} = 1 - [(tc_j + 1) / (df_j \cdot (tfn_{ij}+1))]$$
$$\text{Inf}^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (n_e+0{,}5)] \text{ avec } n_e = n \cdot [1 - [(n-1)/n]^{tc_j}] \quad (5)$$

Enfin, nous avons repris un modèle de langue (LM) [HIE 00] dans lequel les probabilités sont estimées directement en se basant sur les fréquences d'occurrences dans le document d_i ou dans le corpus C. Dans cet article, nous avons repris le modèle de Hiemstra [HIE 00] décrit dans l'équation 6 qui combine une estimation basée sur le document (soit Prob[t_j | d_i]) et sur le corpus (Prob[t_j | C]).

$$\text{Prob}[d_i \mid q] = \text{Prob}[d_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot \text{Prob}[t_j \mid d_i] + (1-\lambda_j) \cdot \text{Prob}[t_j \mid C]] \quad (6)$$
$$\text{avec Prob}[t_j \mid d_i] = tf_{ij} / nt_i \quad \text{et Prob}[t_j \mid C] = df_j / lc \quad \text{avec } lc = \sum_k df_k \quad (7)$$

Dans les formules ci-dessus, λ_j est un facteur de lissage (une constante pour tous les termes t_j, fixée à 0,35) et lc correspond à une estimation de la taille du corpus C.

5. Évaluation

Afin de mesurer la performance [BUC 05], [SAV 06], [VOO 07] de ces divers modèles de recherche d'information, nous avons utilisé l'inverse du rang moyen de la première bonne réponse (MRR ou *Mean Reciprocal Rank*), mesure reflétant le comportement des internautes souhaitant uniquement une seule bonne réponse. De plus, nous pourrons comparer nos mesures de performance avec celles obtenues dans la campagne TREC-5. Afin de savoir si une différence entre deux modèles s'avère statistiquement significative, nous avons opté pour un test bilatéral non-paramétrique (basé sur le ré-échantillonnage aléatoire [SAV 97], avec $\alpha = 5$ %).

Dans un premier temps, nous avons évalué diverses stratégies sans inclure les erreurs de reconnaissance. La section 5.1. présente les performances obtenues en considérant six modèles de recherche, l'emploi de trois enracineurs plus ou moins agressifs et trois représentations différentes des documents. Sur la base de ces résultats, nous pouvons évaluer et analyser l'impact de l'inclusion d'un taux d'erreur en reconnaissance de l'ordre de 5 % à 20 % (section 5.2). Quelques requêtes sont analysées dans la section 5.3. Enfin, dans la dernière section, nous évaluons l'impact d'une procédure d'enrichissement automatique de la requête permettant de remédier partiellement à un taux de reconnaissance imparfait.

5.1. Évaluation des stratégies sur la collection sans erreur

Comme notre mesure d'évaluation se concentre sur la précision, une représentation adéquate correspondrait à un sac de mots, avec l'absence de tout traitement morphologique (évaluation donnée sous la colonne "aucun" dans la table 2). Comme variante, nous pensons que le recours à un enracineur léger [HAR 91] devrait fournir également une performance d'un niveau comparable, voire meilleur. Enfin, l'élimination de suffixes dérivationnels (Porter [POR 80]) risque de nuire quelque peu à la qualité de réponse.

Comme représentation alternative au modèle "sac de mots", nous avons retenu les séquences consécutives de n lettres ou n-grammes, avec une valeur $n = 4$ produisant les meilleurs résultats. Comme autre possibilité, nous suggérons de recourir à une troncature aux n premières lettres (avec $n = 4$), une stratégie que nous pensons adéquate pour une collection de documents bruités. Au niveau des modèles de recherche d'information, nous avons retenu trois modèles probabilistes (Okapi, DFR-PL2, DFR-I(n_e)B2), un modèle de langue (LM) et deux modèles vectoriels (Lnu-ltu et *tf idf*).

Les évaluations indiquées dans la table 2 démontrent que le modèle DFR-I(n_e)B2 propose la meilleure qualité de réponse, quel que soit l'enracineur choisi (colonne "aucun", "léger" ou "Porter") ou selon une représentation par quadri-grammes ou par troncature aux n premières lettres.

	sac de mots			4-gram.	trunc-4
Enracineur	aucun	léger (-'s')	Porter	aucun	aucun
Okapi	0,7620	0,8231	0,8513	0,7730*	0,7826
DFR-PL2	0,7238	0,7458	0,7936	0,6500*	0,6353
DFR-I(n_e)B2	0,7958	0,8343	0,8737	0,8240	0,8126
LM (λ=0,35)	0,7893	0,8077	**0,8807**	0,7029*	0,8209
Lnu-ltu	0,6776	0,7352	0,7703	0,5913	0,6921
tf idf	0,3775*	0,4539*†	0,4214*†	0,3495*	0,3429*
% différence		+ 6,6 %	+ 11,3 %	- 5,7 %	- 1,0 %

***Table 2** : Évaluation de six modèles de recherche d'information selon trois représentations et trois enracineurs (49 requêtes)*

Dans cette table, les différences de performance statistiquement significatives par rapport à la meilleure approche sont notées par un astérisque '*'. Comme on le constate, la performance des autres modèles probabilistes ne s'écarte pas significativement du modèle DFR-I(n_e)B2. La différence de performance avec l'approche *tf idf* s'avère en revanche statistiquement significative.

Si l'on fixe comme référence la performance obtenue en l'absence de tout traitement morphologique (colonne "aucun"), la suppression de la marque du pluriel (colonne "léger") ou l'algorithme de Porter tendent à apporter des performances

supérieures. Toutefois, en moyenne, ces différences de performance sont relativement faibles, soit de + 6,6 % avec un enracineur léger, ou de + 11,3 % avec l'algorithme de Porter. Pour la représentation par 4-grammes, la performance diminue, en moyenne, de quelque -5,7 % ou demeure à un niveau similaire avec la troncature à 4 (-1,0 %). Si l'on applique un test statistique afin de vérifier si ces différences sont réellement significatives, on constate qu'uniquement pour le modèle *tf idf* et avec les enracineurs léger ou Porter, cette différence peut être vue comme significative par rapport à la colonne "aucun" (notée par le symbole '†' dans la table 2).

Enfin, comparé à la meilleure performance obtenue lors de la campagne d'évaluation TREC (MRR = 0,7353) (une comparaison souhaitée par certains auteurs [ARM 09]), les valeurs de la table 2 indiquent clairement que les modèles probabilistes apportent des niveaux supérieurs de performance (e.g., DFR-I(n_e)B2, enracineur léger obtient un MRR de 0,8407, soit une amélioration de 13,5 %.

5.2. Évaluation des collections bruitées

Les performances indiquées dans la table 2 correspondent à l'absence d'erreur de reconnaissance et serviront de référence. Dans la table 3, nous avons repris ces performances sous les colonnes "léger -s" ou "Porter". Avec la présence d'un taux d'erreur de 5 % ou de 20 % en reconnaissance, nous obtenons les valeurs MRR données sous les colonnes "5 %" ou "20 %", avec un enracineur léger ou celui de Porter. La dernière ligne de la table 3 indique qu'en moyenne la qualité diminue de 22 % avec un taux d'erreur de l'ordre de 5 % et d'environ 58 % avec un taux d'erreur de 20 %. On remarquera que l'application d'un enracineur léger ou agressif produit une détérioration relative similaire. Après application de notre test statistique et en prenant comme référence la performance avec le corpus sans erreur, toutes les différences s'avèrent statistiquement significatives.

Enracineur	léger -'s'	5 %	20 %	Porter	5 %	20 %
Okapi	0,8231	0,6181	0,3323	0,8513	0,6113	0,3534
DFR-PL2	0,7458	0,6209	0,3112	0,7936	0,6193	0,3250
DFR-I(n_e)B2	0,8343	0,5899	0,3456	0,8737	0,6157	0,3477
LM (λ=0,35)	0,8077	0,6631	0,3226	0,8807	0,6917	0,3221
Lnu-ltu	0,7352	0,5538	0,2955	0,7703	0,5912	0,3146
tf idf	0,4539	0,3740	0,2511	0,4214	0,3703	0,2255
% différence		-22,3 %	-57,8%		-22,8 %	-58,9 %

***Table 3** : Évaluation selon six modèles, représentation par sac de mots, adjonction de bruit de 5 % ou de 20 % (49 requêtes)*

Cette première analyse se fondait sur une représentation des documents et requêtes par une approche "sac de mots". Dans la table 4, nous avons évalué

l'indexation basée sur des quadri-grammes et la troncature à 4, deux techniques d'indexation devant être plus robustes. Comme pour la table 3, la dégradation de la performance est marquée face à un taux d'erreur de 20 %. Toutefois, la représentation par 4-grammmes résiste mieux qu'une indexation par mots, comme l'indique la dernière ligne de la table 4. Ainsi, la diminution relative de la performance se situe en dessous du seuil de 50 % (précisément de -44,5 %). Toutefois la qualité de la réponse obtenue par une troncature à quatre lettres se dégrade rapidement avec un taux d'erreur élevé. Comme pour la table 3, notre test statistique indique que toutes les différences de performance après l'application de l'OCR s'avèrent statistiquement significatives.

Enracineur	4-gram.	5 %	20 %	trunc-4	5 %	20 %
Okapi	0,7730	0,6472	0,4013	0,7826	0,5698	0,2785
DFR-PL2	0,6500	0,5063	0,3661	0,6353	0,5498	0,3340
DFR-I(n_e)B2	0,8240	**0,6691**	0,4083	0,8126	0,5814	0,3702
LM (λ=0,35)	0,7029	0,5822	**0,4135**	0,8209	0,6174	0,2862
Lnu-ltu	0,5913	0,5128	0,3672	0,6921	0,5009	0,3134
tf· idf	0,3495	0,3029	0,2032	0,3429	0,2817	0,1194
% différence		-17,2 %	-44,5 %		-24,1 %	-58,4 %

***Table 4** : Évaluation selon six modèles, représentation par quadri-grammes, bruit de 5 % ou de 20 % (49 requêtes)*

Lors de la campagne d'évaluation TREC, la performance moyenne sur la collection sans erreur s'élevait à 0,5546, et la moyenne avec un taux d'erreur à 5 % se situait à 0,3401 (soit une perte relative de -38,7 %). En consultant les tables 3 et 4, on remarque que les pertes moyennes se situent aux environs de -20 %. La meilleure qualité des modèles de recherche proposés (Okapi, DFR) explique une dégradation plus faible que celle observée à TREC. En revanche, avec un corpus renfermant un taux d'erreur de 20 %, la performance moyenne de TREC se situait à 0,2663, soit une perte relative de -52 %. Nos résultats indiquent des niveaux de détérioration similaires.

Selon les tables 3 et 4, la meilleure performance s'obtient en recourant à une indexation par 4-grammes et en utilisant le modèle probabiliste DFR-I(n_e)B2 ou LM. Comparé au meilleur résultat de TREC, cette approche améliore la performance de 17 % (0,6691 vs. 0,5737) avec un taux d'erreur de 5 % (différence statistiquement significative).

5.3. Analyse de quelques requêtes

Afin de mieux comprendre les effets des différentes stratégies d'indexation et de recherche face aux erreurs OCR, nous avons analysé quelques requêtes. Pour comprendre la différence entre une représentation par mots ou des *n*-grammes,

l'interrogation n° 36 “headband” constitue un bon exemple. Avec l'emploi de mots et quel que soit l'enracineur utilisé, le système de recherche place la bonne réponse en première position. Lorsque cette requête est représentée par des quadrigrammes, la séquence {“head”, “eadb”, “adba”, “dban” et “band”} est générée. Sachant que le seul article pertinent possède une seule fois le terme “headband”, ce document rentre en conflit avec tous les autres articles possédant plusieurs occurrences des 4-grammes “head” ou “band” (e.g., comme avec les mots “broadband”, “baseband”, etc.). La bonne réponse apparaît en 85 position, retournant une très faible performance pour cette interrogation.

Pour d'autres demandes, la représentation par quadri-grammes s'avère plus avantageuse comme pour la requête n° 13 “I am looking for theft data on the Chevrolet Corsica”. La bonne réponse contient quinze fois le nom propre “chevrolet”, une fois les mots “theft”, “thefts” et “corsica”, mais pas le terme “data”. Sur l'ensemble de la collection, ces termes apparaissaient dans 36 documents pour “chevrolet”, 254 pour “theft”, 16 pour “thefts” et un seul article pour “corsica”. Que l'on recherche dans la collection originale ou dans celle avec 5 % ou 20 % d'erreur, la bonne réponse apparaît toujours en première position. Les termes apparaissent dans une orthographe différente comme “che.vrolet” “chevrotet”, “fwhevrolet” ou “L-orsica”. Avec une représentation par 4-grammes, on rencontre plusieurs appariements entre la requête et la bonne réponse comme, par exemple, pour les termes “evro” “vrol” tirés de “chevrolet” ou “rsic” de “corsica”. Ces exemples illustrent les différences existantes entre les représentations, en soulignant que parfois une forme d'indexation s'avère meilleure qu'une autre.

5.4. Pseudo-rétroaction positive

Il est reconnu que le recours à une pseudo-rétroaction [BUC 96] afin d’élargir automatiquement les requêtes courtes, permet d’augmenter la qualité de la recherche d'information. Une telle approche semble, *a priori*, aussi attractive dans notre contexte puisque le système a l'opportunité d'extraire des termes correctement orthographiés des documents classés en tête de liste.

Modèle	4-grammes	5 %	20 %
DFR-I(n_e)B2	0,8240	0,6691	0,4083
3 docs / 10 termes	0,4341†	0,3265†	0,2563†
3 docs / 20 termes	0,4193†	0,3438†	0,2449†
3 docs / 50 termes	0,5071†	0,3696†	0,2667
5 docs / 10 termes	0,2697†	0,2280†	0,1607†
5 docs / 20 termes	0,3183†	0,2419†	0,1594†
5 docs / 50 termes	0,3415†	0,2615†	0,1629†

***Table 5** : Évaluation de la pseudo-rétroaction positive, représentation par quadri-grammes, bruit de 5 % ou de 20 % (49 requêtes)*

Afin de procéder à une expansion automatique, nous avons implémenté l'approche de Rocchio [ROC 71] avec les constantes $\alpha = 0{,}75$ et $\beta = 0{,}75$ et en incluant entre 10 et 50 nouveaux *n*-grammes extraits des 3 ou 5 premiers articles dépistés. Les résultats obtenus avec les trois versions de notre collection sont indiqués dans la table 5. Ces performances sont décevantes, indiquant des diminutions dans tous les cas de figure, que la collection comporte ou non des erreurs. De plus, ces différences sont statistiquement toujours significatives par rapport à la performance obtenue avant la rétroaction (indiquées par le symbole '†').

6. Conclusion

Dans cette étude, nous avons analysé la dégradation de la performance en recherche face à un corpus contenant un taux d'erreur en reconnaissance des caractères de l'ordre de 5 % ou 20 %. Face à six modèles de recherche, trois représentations des documents et trois enracineurs, les tables 3 et 4 indiquent clairement une diminution relative d'environ 20 % face à un taux d'erreur de 5 %. Si le taux d'erreur s'élève à 20 %, cette dégradation s'élève entre -47 % et -65 %. Au niveau statistique, ces différences de performances sont toujours significatives. Dans ces évaluations, le modèle probabiliste *Divergence from Randomness* ou un modèle de langue apportent la meilleure performance.

Afin de réduire cette érosion de la qualité, nous avons envisagé d'ignorer la suppression de certains suffixes. Cette solution apporte des résultats similaires à ceux obtenus après la suppression des -'s' finaux. En revanche, la représentation des documents et des requêtes par des séquences de *n*-grammes permet de corriger quelque peu la perte de la qualité de la recherche. Par contre, si le taux d'erreur s'élève à 20 %, même cette forme d'indexation donne des performances moyennes dont la diminution s'avère proche des 50 %.

Pour compenser cette perte, le recours à un enrichissement automatique par pseudo-rétroaction n'apporte pas d'amélioration sensible du rang de la première bonne réponse (voir table 5). L'emploi de cette technique dégrade également les performances en l'absence d'erreur, démontrant ainsi que cette approche ne devrait pas être appliquée lorsque l'on recherche un nombre restreint de bonnes réponses.

Les résultats de cette analyse indiquent qu'en présence d'un taux d'erreur en reconnaissance de l'ordre de 5 %, les techniques habituelles de recherche d'information continuent à offrir une qualité de réponse acceptable (perte relative de qualité de l'ordre de 20 %). La représentation par *n*-grammes semble dans ce cas apporter une meilleure performance qu'une indexation par mots. Si le taux d'erreur augmente pour se situer à environ 20 %, les techniques usuelles (enracineur léger ou non, pseudo-rétroaction positive) ne suffisent plus à contrer l'érosion de la qualité.

Remerciements

Cette recherche a été financée en partie par le Fonds national suisse pour la recherche scientifique (subside CRSI22_125220).

7. Bibliographie

[ABD 07] Abdou, S., & Savoy, J. "Considérations sur l'évaluation de la robustesse en recherche d'information", Actes CORIA'07, St-Etienne, 2007, p. 5-30.

[AMA 02] Amati, G., & van Rijsbergen, C.J. "Probabilistic models of information retrieval based on measuring the divergence from randomness", ACM Transactions on Information Systems, vol. 20, nº 4, 2002, p. 357-389.

[ARM 09] Armstrong, T.G., Moffat, A., Webber, W. & Zobel, J. "Improvements that don't add up: Ad hoc retrieval results since 1998" Proceedings ACM-CIKM, Hong Kong, 2009, p. 601-609.

[BAL 97] Ballerini, J.P., Büchel, M., Domering, R., Knaus, D., Mateev, B., Mittendorf, E., Schäuble, P., Sheridan, P., & Wechsler, M. "SPIDER retrieval system at TREC-5", Proceedings of TREC-5, NIST Publication #500-238, Gaithersburg, 1997, p. 217-228.

[BUC 96] Buckley, C., Singhal, A., Mitra, M., & Salton, G. "New retrieval approaches using SMART", Proceedings of TREC-4, NIST Publication #500-236, Gaithersburg (MD), 1996, p. 25-48.

[BUC 05] Buckley, C., & Voorhees, E. "Retrieval system evaluation". In E.M. Voorhees & D.K. Harman (Eds), "TREC, Experiment and Evaluation in Information Retrieval", The MIT Press, Cambridge (MA), 2005, p. 53-75.

[CAL 02] Callan, J., Kantor, P., & Grossman, D. "Informtion retrieval and OCR: From converting content to grasping meaning", SIGIR Forum, vol. 36, nº 2, 2002, p. 58-61.

[CUC 04] Cucerzan, S., & Brill, E. "Spelling correction as an iterative process that exploits the collective knowledge of the web users", Proceedings EMNLP, 2004, p. 293-300.

[DEC 09] De Closets, F. "Zéro faute. L'orthographe une passion française", Mille et une nuits, Paris, 2009.

[DOL 10] Dolamic, J. & Savoy, J. "Comparative study of indexing and search strategies for the Hindi, Marathi and Bengali languages", ACM Transactions on Asian Information Languages, vol. 9, nº 3, 2010, p.1-24.

[GOV 09] Govindaraju, V., Cao, H., & Bhardwaj, A. "Handwritten document retrieval strategies", Proceedings AND'09 Workshop on Analytics for Noisy Unstructured Text Data, 2009, p. 3-7.

[HAR 91] Harman, D. "How effective is suffixing?", Journal of the American Society for Information Science, vol. 42, nº 1, 1991, p. 7-15.

[HAR 97] Harding, S.M, Croft, W.B., & Weir, C. "Probabilistic retrieval of OCR degraded text using *n*-grams", Proceedings of the ECDL'97, Berlin, 1997, p. 345-359.

[HIE 00] Hiemstra, D. "Using language models for information retrieval", CTIT Ph.D. Thesis, 2000.

[KAN 00] Kantor, P.B., & Voorhees, E.M. "The TREC-5 confusion track: Comparing retrieval methods for scanned text", IR Journal, vol. 2, nº 2-3, 2000, p. 165-176.

[MAN 08] Manning, C., Raghavan, P., & Schütze, H. "Introduction to Information Retrieval", Cambridge University Press, Cambridge (UK), 2008.

[MCN 04] McNamee, P., & Mayfield, J. "Character *n*-gram tokenization for European language text retrieval", IR Journal, vol. 7, nº 1-2, 2004, p. 73-97.

[MIC 00] Mitra, M., & Chaudhuri, B.B. "Information retrieval from documents: A survey", IR Journal, vol. 2, nº 2-3, 2000, p. 141-163.

[MIS 00] Mittendorf, E., & Schäuble, P. "Information retrieval can cope with many errors", IR Journal, vol. 3, nº 3, 2000, p. 189-216.

[PIL 06] Pilz, T., Luther, W., Fuhr, N., & Ammon, U. "Rule-based search in text databases with nonstandard orthography", Literacy & Linguistic Computing, vol. 21, nº 2, 2006, p. 179-186.

[POR 80] Porter, M.F. "An algorithm for suffix stripping", Program, 14, 1980, p. 130-137.

[ROB 00] Robertson, S.E., Walker, S., & Beaulieu, M. "Experimentation as a way of life: Okapi at TREC", Information Processing & Management, vol. 36, nº 1, 2000, p. 95-108.

[ROC 71] Rocchio, J.J.Jr. "Relevance feedback in information retrieval", In G. Salton (Ed.), The SMART Retrieval System. Prentice-Hall Inc., Englewood Cliffs, 1971, p. 313-323

[SAV 97] Savoy, J. "Statistical inference in retrieval effectiveness evaluation", Information Processing & Management, vol. 33, nº 4, 1997, p. 495-512.

[SAV 06] Savoy, J. "Un regard statistique sur l'évaluation de performance : L'exemple de CLEF 2005", Actes CORIA'06, Lyon, 2006, p. 73-84.

[SAV 08] Savoy, J., & Dolamic, L. (2008). "Variations autour de *tf idf* et du moteur Lucene", Actes Analyse statistique des Données Textuelles, 2008, p. 1047-1058.

[SAV 10] Savoy, J. "Représentation comparative: Applications au discours électoral en Suisse, France et États-Unis", Document Numérique, vol. 13, nº 1, 2010, p. 111-135.

[TAG 94] Tagva, K., Borsack, J., & Condit, A. "Results of applying probabilistic IR to OCR text", Proceedings of the ACM-SIGIR, Dublin, 1994, p. 202-211.

[TAG 96] Tagva, K., Borsack, J., & Condit, A. "Evaluation of model-based retrieval effectiveness with OCR text", ACM Transactions on Information Systems, vol. 14, nº 1, 1996, p. 64-93.

[TAG 04] Tagva, K., Nartker, T., & Borsack, J. "Information access in presence of OCR errors", Proceedings of the ACM Workshop on Hardcopy Document Processing, Washington (DC), 2004, p. 1-8.

[VOO 05] Voorhees, E.M., & Garofolo, J.S. "Retrieving noisy text", In E.M. Voorhees & D.K. Harman (Eds), "TREC, Experiment and Evaluation in Information Retrieval", The MIT Press, Cambridge (MA), 2005, p. 183-197.

[VOO 07] Voorhees, E.M. "TREC: Continuing information retrieval's tradition of experimentation", Communications of the ACM, vol. 50, nº 11, 2007, p. 51-54.