Université de Neuchâtel

Recherche d'Information Plurilingue

par

Samir Abdou

Thèse

présentée à la faculté des sciences
pour l'obtention du grade de Docteur ès Sciences

Composition du Jury

Prof. Jacques Savoy (Directeur de thèse)
Université de Neuchâtel, Suisse

Prof. Patrice Bellot
Université d'Avignon, France

Prof. Pascal Felber
Université de Neuchâtel, Suisse

Prof. Rolf Ingold
Université de Fribourg, Suisse

Juin 2007

Université de Neuchâtel unine

FACULTE DES SCIENCES
Secrétariat-Décanat de la faculté
- Rue Emile-Argand 11
- CP 158
- CH-2009 Neuchâtel

IMPRIMATUR POUR LA THESE

Recherche d'information plurilingue

Samir ABDOU

UNIVERSITE DE NEUCHATEL

FACULTE DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel,
sur le rapport des membres du jury

MM. J. Savoy (directeur de thèse),
P. Felber, R. Ingold (Fribourg)
et P. Bellot (Avignon F)

autorise l'impression de la présente thèse.

Neuchâtel, le 26 juillet 2007

Le doyen :
T. Ward

UNIVERSITE DE NEUCHATEL
FACULTE DES SCIENCES
Secrétariat-décanat de la faculté
Rue Emile-Argand 11 - CP 158
CH-2009 Neuchâtel

RÉSUMÉ.

Dans cette thèse, nous présentons nos investigations en recherche d'information selon deux contextes. Premièrement, nous nous sommes intéressés à l'aspect plurilingue de la Toile en abordant le développement de moteurs de recherche pour des langues présentant des caractéristiques visuelles, morphologiques et syntaxiques très différentes des langues indo-européennes. Plus précisément, nous avons proposé des stratégies de recherche pour les langues chinoise (traditionnelle), japonaise et coréenne, ainsi que pour la langue anglaise à des fins de comparaison. A cet effet, nous avons utilisé des corpus de dépêches d'agences représentant en général un contexte proche de la réalité du Web puisque la consultation de nouvelles constitue l'une des raisons importantes de la navigation sur Internet. Nous avons réalisé cette partie du travail avec le but de fournir la meilleure qualité de recherche possible pour chacune de ces langues. Plus concrètement, nous avons comparé divers modèles de recherche ainsi que diverses stratégies d'indexation. Nous avons également proposé une nouvelle approche de pseudo-rétroaction permettant d'améliorer significativement la performance de nos moteurs de recherche. Ces expériences nous ont ainsi permis de démontrer qu'une qualité de recherche optimale ne peut être obtenue qu'en considérant les particularités de chacune de ces langues.

Deuxièmement, nous avons abordé le développement de moteurs de recherche dans le contexte spécifique de la biomédecine. A cet effet, nous avons utilisé un corpus de notices bibliographiques rédigées en anglais et issues du domaine biomédical (un tiers du corpus MEDLINE). En adaptant nos stratégies développées pour la langue anglaise, nous avons d'abord comparé divers modèles de recherche. Ensuite, nous avons évalué l'impact des descripteurs manuellement attribués sur la qualité de la recherche. Enfin, une comparaison de notre approche de pseudo-rétroaction avec celle de Rocchio a été réalisée. Dans ce contexte, ces expériences ont démontré que l'indexation manuelle permet d'améliorer très nettement la performance, et ce quel que soit le modèle de recherche. L'évaluation de la rétroaction a donné des résultats contradictoires. Tandis que l'approche de Rocchio a sérieusement détérioré la performance du système, notre stratégie de pseudo-rétroaction a permis de l'améliorer.

Finalement, nous avons porté un regard critique sur quelques facettes liées à l'évaluation en recherche d'information. Nous avons comparé diverses métriques permettant d'évaluer divers critères de la qualité de recherche. Nous avons réalisé cette expérience sur deux applications, à savoir le classement des modèles de recherche et la pseudo-rétroaction selon Rocchio. Ces évaluations ont permis d'identifier quelques circonstances dans lesquelles nos diverses approches de recherche pourraient rencontrer des difficultés. Nous avons ainsi montré l'importance de considérer diverses métriques avec leurs avantages et inconvénients pour évaluer l'efficacité des systèmes de recherche d'information.

MOTS-CLES : recherche d'information plurilingue, langues asiatiques, domaine spécifique, indexation, moteur de recherche, pseudo-rétroaction

KEYWORDS: plurilingual information retrieval, Asian languages, domain-specific, indexing, search engine, pseudo-relevance feedback

REMERCIEMENTS

De nombreuses personnes ont contribué de près ou de loin au succès de ce travail de thèse.

Je remercie mes parents, mes sœurs et mes frères pour leur aide et encouragement durant toutes mes études. Je remercie particulièrement mon frère Farid et ma mère Houria pour leur soutien à tous les niveaux. Je leur serai toujours redevable de tous les efforts qu'ils ont fournis à mon égard.

Je tiens également à vivement remercier mon amie Sandra pour sa patience, sa gentillesse et son encouragement durant toute cette période.

J'aimerais exprimer ma gratitude et mes sincères remerciements à mon directeur de thèse, le Professeur Jacques Savoy, qui m'a offert l'opportunité de travailler avec lui et de découvrir le monde de la recherche d'information. Son enthousiasme, son encadrement ainsi que son humanisme dans le travail ont constitué un environnement idéal pour mener à bien ce projet.

Cette thèse a été évaluée par le jury composé des professeurs Jacques Savoy, Pascal Felber (Université de Neuchâtel), Patrice Bellot (Université d'Avignon) et Rolf Ingold (Université de Fribourg). Je les remercie tous pour le temps qu'ils ont consacré à l'expertise de cette thèse.

Enfin, cette recherche a été en partie soutenue par le Fonds National Suisse de la Recherche Scientifique avec les subsides n° 200020-103420 et n° 200020-115866. L'université de Neuchâtel a également mis à ma disposition un poste partiel d'assistant.

TABLE DES MATIÈRES

1. Introduction

1.1. *Motivations*

La recherche d'information (RI) connait une popularité grandissante depuis la généralisation de l'Internet et des ordinateurs personnels. Alors que les systèmes de stockage et de dépistage d'information étaient essentiellement réservés à des professionnels de domaines spécialisés comme le droit, l'aéronautique, la recherche médicale ou les organisations gouvernementales, l'évolution très rapide de l'Internet et la mise à disposition d'un savoir planétaire a permis de confronter le domaine de la RI à l'émergence de multiples défis (Allan *et al.*, 2002). Dans le cadre de cette thèse, nous avons porté nos efforts d'investigations sur deux aspects en particulier.

Le multilinguisme du Web

Outre la croissance du volume d'informations mis à disposition ou la présence de multiples supports (du texte ASCII aux pages XML, de la photographie aux séquences musicales ou vidéo), le caractère plurilingue de la Toile (le Web) représente à nos yeux un enjeu considérable. Dans ce contexte, l'importance croissante de langues autres que l'anglais a suscité le développement d'outils et de techniques automatiques afin de permettre des traitements informatiques appropriés dans ces diverses langues. Ce besoin n'est pas marginal et quelques chiffres peuvent résumer son importance.

L'Internet à ses débuts était pratiquement anglophone, et ce jusqu'à la fin des années 1990. En effet, le pourcentage de pages Web rédigées en anglais s'élevait à 75 % en septembre 1998 avant de chuter à 45 % en mars 2005[1]. Par ailleurs, en décembre 2000[2], la population d'internautes naviguant en anglais s'élevait à 47 % pour environ 407 millions d'internautes. Par contre, en mars 2007, avec une population de 1,1 milliard d'individus, cette proportion était estimée à 29,5 % contre 14,3 % pour le chinois, 8 % pour l'espagnol, 7,7 % pour le japonais et 5,3 % pour l'allemand. La langue française arrive en sixième position avec 5 %, suivi du portugais avec 3,6 % et du coréen avec 3,1 %. A côté de ces valeurs, il convient de tenir compte du dynamisme de l'évolution linguistique. Ainsi, durant la période de 1997 à 2002, Wei (2004) a montré que le nombre d'utilisateurs en ligne par langue a progressé, en moyenne et par année, de 165 % pour le chinois, 140 % pour le coréen, 59 % pour le japonais et seulement 28 % pour l'anglais. Sur cette base, on estime que l'utilisation des langues asiatiques sur le Web va atteindre des valeurs comparables voire supérieures à celle de l'anglais.

En plus de ces estimations, d'autres facettes liées au plurilinguisme du Web doivent être considérées. En effet, certains usagers parlent ou, pour le moins,

---

[1] Voir le site http://funredes.org

[2] Voir le site http://www.internetworldstats.com

possèdent des connaissances de plusieurs langues. Ainsi, dans les pays bilingues à l'image du Canada ou de la Finlande, ou multilingues comme la Suisse, le Singapour, l'Inde ou la Communauté Européenne, l'accès électronique aisé et efficace à l'information indépendamment de la langue de l'usager devient un enjeu important. Par exemple, un juriste travaillant sur le droit européen doit pouvoir, en réponse à une requête rédigée dans une langue, dépister aisément des textes légaux, que ceux-ci soient écrits en italien, anglais, allemand ou finnois. Avec la globalisation, les gestionnaires d'entreprises multinationales ou d'organisations internationales à l'image de l'OMC doivent parfois faire face à des situations similaires. Par ailleurs, certains utilisateurs peuvent rencontrer des difficultés à exprimer leur besoin d'information dans une langue étrangère bien qu'ils puissent comprendre des documents rédigés dans cette langue (Oard & Resnik, 1999). Dans ces diverses situations, les stratégies de recherche monolingues ne répondent pas à toutes les demandes et quelquefois l'usager souhaiterait que le système de RI puisse l'aider à franchir la barrière linguistique (Grefenstette, 1998).

Pour répondre à ces différentes exigences, la mise en place de systèmes monolingues permettant de dépister efficacement l'information constitue une étape fondamentale pour concevoir des systèmes de RI translinguistiques (Braschler & Peters, 2004). Alors que la langue anglaise a dominé le développement et le perfectionnement de tels systèmes depuis le début du domaine de la RI, l'intérêt pour d'autres langues n'a réellement commencé qu'avec l'initiative de TREC[3] (Voorhees & Harman, 2005) qui a suscité le besoin de renforcer les efforts pour la mise en œuvre de moteurs de recherche pour d'autres langues que l'anglais (Harman, 2005). Mais ce n'est qu'à partir de 1994 que cet intérêt se concrétise avec l'introduction de la langue espagnole dans la tâche *ad hoc* de TREC-3, du chinois simplifié en 1996 (TREC-5), de l'allemand, français et italien en 1997 (TREC-6) dans la tâche multilingue, du chinois traditionnel en 2000 (TREC-9) et, enfin, de la langue arabe en 2001 (TREC-10). Cet engouement pour le traitement d'autres langues s'est ensuite intensifié et concentré avec les campagnes d'évaluation de NTCIR[4] (Kando & Adachi, 2004 ; Kishida *et al.*, 2004 ; Nakagawa *et al.*, 2005) depuis 1999 pour les langues asiatiques (japonais, chinois et coréen) et CLEF[5] (Braschler & Peters, 2004 ; Peters *et al.*, 2005 ; Peters *et al.*, 2006) depuis 2000 pour les langues européennes (français, néerlandais, italien, espagnol, allemand, suédois, russe, finnois, bulgare, hongrois et portugais-brésilien).

Les systèmes de RI dédiés et contextuels

La Toile n'est qu'un aspect de l'éventail des systèmes d'informations. La partie invisible du Web, inaccessible via des moteurs de recherche conventionnels, renferme une masse de données gigantesque. Selon les estimations de Bergman

---

[3] *Text REtrieval Conference* : http://trec.nist.gov

[4] *NII Test Collection for IR Systems* : http://research.nii.ac.jp/ntcir/index-en.html

[5] *Cross Language Evaluation Forum* : http://www.clef-campaign.org

(2001), celle-ci est 500 fois plus importante que la partie visible. L'auteur affirme également que la qualité des informations et ressources du Web caché est souvent meilleure que celle de la partie visible. Dans le but de rendre accessible ces connaissances, de nombreux moteurs de recherche spécialisés proposent l'accès à des fonds documentaires dans divers contextes comme les bibliothèques numériques, les quotidiens et autres journaux scientifiques, les documents et archives publiques (par exemple, la radio ou la télévision) ainsi que les lois, règlements et décisions des tribunaux. Par exemple MEDLINE (*Medical Literature Analysis and Retrieval System Online*), accessible au travers de l'interface PUBMED[6], constitue la principale source d'informations pour les biologistes et les spécialistes du domaine médical. Le nombre de demandes soumises à ce système s'élevait à 244 millions pour l'année 2000 ; ce chiffre est passé à 896 millions de requêtes en 2006, soit une augmentation de 296 %. Durant cette même période, le nombre de notices bibliographiques indexées a progressé d'environ 25 % pour s'établir à quelques 14 millions de citations en septembre 2006. Pour le domaine du légal, les systèmes commerciaux LEXIS-NEXIS et WESTLAW[7] se partagent le marché des fournisseurs d'accès à des banques documentaires de la jurisprudence.

Dans les domaines spécifiques, l'information est généralement plus pointue et possède une terminologie précise (voir un exemple d'une notice bibliographique dans l'annexe E) comparée au contenu des dépêches d'agences (voir un exemple de nouvelle de presse dans l'annexe D), qui couvrent souvent des sujets généraux et variés de l'actualité. De plus, les besoins des utilisateurs dans ces deux contextes sont différents. Premièrement, un utilisateur moyen sur le Web peut être satisfait des dix ou vingt premières réponses du système car il désire un nombre très limité, voire une seule bonne réponse[8]. Par contre, un spécialiste dans le domaine légal ou médical souhaiterait, en revanche, obtenir une liste plus exhaustive de réponses pertinentes. Autrement dit, retrouver tous les documents pertinents (niveau de rappel élevé) avec un minimum de bruit (haute précision). Une telle exigence s'avère difficile à satisfaire. Deuxièmement, la formulation des besoins est plus précise pour un expert d'un domaine particulier que pour un usager moyen. Sur le Web, les requêtes comprennent en moyenne 2,2 termes tandis que les requêtes soumises à des systèmes commerciaux comptent 14,8 termes en moyenne (Jansen *et al.*, 2000).

D'autre part, le contenu spécialisé peut révéler de nouveaux phénomènes. De nombreuses causes peuvent entraver l'appariement entre la requête et les documents. Ainsi dans le domaine biomédical, où la littérature est très prolifique, l'absence de convention unanimement utilisée dans l'appellation des gènes, protéines et autres noms est source d'ambigüités qui engendrent des problèmes importants pour le

---

6 Voir le site http://www.ncbi.nlm.nih.gov/entrez/

7 Voir les sites : http://www.westlaw.com et http://www.lexisnexis.com

8 Selon une récente étude (Jansen & Spink, 2006), le pourcentage d'internautes consultant uniquement la première page fournie par le moteur de recherche s'élevait à 73 % en 2002 contre seulement 29 % en 1997.

dépistage de l'information. En particulier et contrairement aux dépêches d'agences, les phénomènes linguistiques tels que la synonymie et l'homonymie sont très fréquents dans la littérature biomédicale. Ceci est essentiellement dû à la dynamique et au développement de multiples domaines sous-jacents et souvent peu connexes. Par exemple, les gènes WAF1 et CIP1 ont été découverts en 1993 indépendamment par deux équipes de recherche travaillant sur le même sujet ; ces deux gènes se sont par la suite révélés identiques, et sont devenus par conséquent synonymes parmi d'autres (p21, SDI1, CAP20 ou MDA-60). L'homonymie relève d'une problématique similaire, spécialement avec les acronymes ou les abréviations (Yu *et al.*, 2006). Par exemple, le symbole PSA peut se référer aux gènes « *prostate specific antigen* », « *puromycin-sensitive aminopeptidase* » ou « *phosphoserine aminotransferase* » ; mais peut également décrire des protéines ou autres concepts comme « *psoriasis arthritis* », « *pig serum albumin* » ou « *poultry science administration* ». En plus de ces définitions, le symbole PSA possède également plus d'une centaine de significations et de variantes différentes (Weeber *et al.*, 2003).

Enfin, la variation orthographique constitue un autre problème pouvant rendre difficile la correspondance entre la requête et les documents. Parmi les causes à l'origine de cet écueil, nous pouvons citer les erreurs d'orthographe, la ponctuation et la typographie alternatives, la transcription de noms étrangers ou les variations linguistiques (dialectes). Dans ce dernier cas et pour la langue anglaise par exemple, certains mots possèdent des formes différentes selon la région linguistique, à savoir britannique ou américaine comme pour « *colour vs. color* » ou « *programme vs. program* ». Quant aux noms propres, et en particulier ceux d'origine étrangère ou peu fréquents, leur transcription dans une autre langue tend à produire diverses formes possibles (« Gorbachev », « Gorbacheff » ou « Gorbachov » pour n'en citer que trois). Dans le domaine biomédical, ce phénomène prend une ampleur considérable avec les noms de maladies, de gènes et de protéines. Par exemple, la protéine « NF-kappaB » peut apparaître sous les formes « NF-kappa B » ou « NF-kappa-B ».

Afin de faire face à ces différents problèmes, de nombreuses ressources ont été développées comme des thesaurus, dictionnaires et autres listes ou bases de données. Les applications liées à l'utilisation de ces ressources en RI sont multiples : par exemple, la correction et la génération de variantes orthographiques, l'indexation à l'aide d'un vocabulaire contrôlé, l'expansion automatique de requêtes ou l'assistance de l'usager dans la formulation des requêtes. Cependant, les résultats obtenus avec l'emploi de ces ressources sont quelque peu lacunaires et quelquefois contradictoires. Dans le but d'améliorer les techniques de la RI et proposer de nouvelles stratégies permettant de répondre aux particularités des domaines spécifiques, de nombreuses tâches ont été introduites dans les conférences TREC (*Genomics* pour le domaine biomédical), NTCIR (*Patent* pour la recherche dans un corpus de brevets) et CLEF (*GIRT* pour le domaine des sciences sociales).

1.2. *Problématique et objectifs*

Dans le cadre de cette dissertation et au regard des motivations présentées dans la section précédente, nos travaux de recherche portaient essentiellement sur les trois questions suivantes :

1. Est-il possible de concevoir un moteur de recherche générique couvrant toutes les langues écrites et offrant les meilleures performances dans chacune de ces langues ?

2. En se limitant à un domaine particulier du savoir, comment bénéficier de ce domaine pour améliorer le(s) moteur(s) de recherche de la première question ?

3. Pourquoi et dans quelles circonstances le(s) moteur(s) de recherche de la première et deuxième question peuvent rencontrer des difficultés face à certaines requêtes ?

Pour la mise en œuvre d'un moteur de recherche monolingue, Harman (1995 ; 2005) estime que la simple adaptation des techniques développées pour la langue anglaise, moyennant une liste de mots-outils et un enracineur, permet d'atteindre des performances jugées acceptables. Cependant, la variabilité linguistique en RI touche de nombreux aspects comme la morphologie[9], les accents et autres diacritiques, la syntaxe, la liste de mots-outils, le codage ou encore le style, qui peut varier pour une même langue d'un contexte (l'exemple des dépêches de presse, voir l'annexe D) à un autre (l'exemple des articles scientifiques, voir l'annexe E).

En prenant ce dernier cas, un moteur de recherche performant pour dépister un besoin d'information dans une collection d'articles de presse (annexe D) ne satisfera qu'en partie les exigences d'un spécialiste recherchant des informations dans un corpus d'articles scientifiques du domaine biomédical (annexe E). En effet, compte tenu des propriétés et difficultés relatives aux domaines spécialisés que nous avons abordées précédemment, la simple adaptation des outils généralistes de la langue anglaise ne suffira certainement pas pour atteindre des performances optimales, voire satisfaisantes. De plus dans ce cas précis, comme il s'agit d'articles rédigés en langue anglaise mais dans deux contextes distincts (style journalistique *vs.* biomédical), aucune adaptation ne serait requise, tout au plus une liste adéquate de mots-outils pourrait être différente lors de l'indexation des documents et des requêtes. Par ailleurs, une étude comparant diverses stratégies d'enracinement sur la

---

[9] La morphologie est la branche de la linguistique qui étudie la structure et la formation des mots. On distingue généralement trois processus, soit la *morphologie flexionnelle* (genre, cas et nombre pour les noms et adjectifs, et mode, temps et personne pour les verbes), la *morphologie dérivationnelle* (dérivation des mots à l'aide de divers affixes) et la *morphologie compositionnelle* (composition de nouveaux mots à partir d'autres mots).

collection MEDLINE (corpus d'articles biomédicaux) révèle que celles-ci n'ont qu'une influence marginale sur la performance des systèmes (Abdou *et al.*, 2006).

Concernant la morphologie, Pirkola (2001) a démontré que la variabilité des propriétés morphologiques entre les langues est élevée. D'une part, certaines langues comme l'italien, l'espagnol ou le français présentent de nombreuses similitudes avec la langue anglaise. Dans ces cas et d'un point de vue linguistique, l'anglais constitue un point de départ intéressant pour aborder ces langues romanes en RI (Jones, 2005). Certes, l'usage d'une liste de mots très fréquents et d'un enracineur, pour le moins léger, peut fournir d'intéressants résultats (Savoy, 2002). Mais ces bonnes performances ne sont pas uniquement dues à l'usage de ces deux outils. D'autres facteurs pouvant contribuer à une recherche efficace ne doivent pas être négligés, même pour des langues a priori similaires : par exemple, le rôle et l'importance des accents pour ces langues n'est pas le même. Ainsi, Hollink *et al.* (2004)[10] ont montré que la substitution des lettres accentuées par leurs formes non accentuées ne produit pas d'effet sur les langues anglaise ou allemande (environ + 2 % de précision moyenne). En revanche, cette normalisation s'est avérée fructueuse et statistiquement significative pour d'autres langues avec une amélioration de la précision moyenne d'environ 8 % pour l'italien, 10 % pour le hollandais, 19 % pour le français et 20 % pour le suédois.

D'autre part, les langues comme l'allemand, le suédois, le finnois ou le hongrois, présentent des caractéristiques différentes au regard de la langue anglaise. En plus de leur richesse en termes de morphologie flexionnelle et dérivationnelle, ces langues se caractérisent par l'usage fréquent de la composition pour la formation de nouveaux mots par concaténation. Dans la langue allemande par exemple, le mot « Computersicherheit » est composé des mots « Computer » et « Sicherheit ». Un système ignorant cette particularité rencontrera des difficultés à dépister des documents contenant uniquement la forme composée lorsque la requête ne contient que les composants. Pour ces langues, la suppression des mots très fréquents et l'application d'un enracineur ne permettent pas toujours d'atteindre les meilleures performances. Certes, ces outils améliorent la qualité de la recherche pour certaines langues par rapport à une approche ignorant ces traitements. Mais ces opérations demeurent inefficaces pour d'autres langues.

Ainsi pour le suédois, Hollink *et al.* (2004) ont observé une amélioration marginale d'environ 2 % par rapport à une approche ignorant l'enracinement tandis que pour les langues allemande et finnoise, une amélioration significative d'environ 7 % respectivement 30 % a été observée. Pourtant, ces performances ne sont pas optimales. En effet, face au phénomène des mots composés de ces langues, les systèmes peinent à dépister de nombreux documents pertinents. A cet effet,

[10] Les travaux de Hollink *et al.* (2004) ont été réalisés en utilisant les collections-tests de CLEF 2002 et le modèle vectoriel « Lnu » (Singhal *et al.*, 1996). Ces notions de modèle vectoriel et de collection-test sont définies respectivement dans les sections 1.4 et 1.5.

plusieurs auteurs suggèrent de procéder à la décomposition de ces mots (Savoy, 2003 ; Chen, 2003 ; Braschler & Ripplinger, 2004). En appliquant cette stratégie après enracinement, Hollink *et al.* (2004) ont amélioré la performance de leurs systèmes par rapport à une approche par enracinement d'environ 10 % pour la langue finnoise, 25 % pour le suédois et 16 % pour l'allemand. Pour cette dernière, Braschler et Ripplinger (2004) ont montré que la décomposition contribue nettement plus que l'enracinement à l'amélioration de la précision moyenne, soit 16 % à 34 % sur des requêtes courtes et 9 % à 28 % sur des requêtes longues. Sur la langue hongroise, la décomposition des mots génère une amélioration de la qualité de recherche d'environ 8 % par rapport à une approche par enracinement (Savoy & Abdou, 2006).

Une approche permettant d'éviter l'utilisation de méthodes nécessitant des connaissances linguistiques est de recourir à la segmentation des mots en *n*-grammes. L'idée de cette stratégie est de parcourir le mot ou le texte avec une fenêtre de taille *n* caractères en la déplaçant caractère par caractère pour obtenir les unités d'indexation. Par exemple le mot « recherche », en utilisant une fenêtre de taille 5, génèrera l'ensemble suivant d'unités d'indexation : { « reche », « echer », « cherc », « herch » et « erche »}. Le choix optimal de la taille des *n*-grammes semble dépendant de la langue, du modèle de recherche et de la méthode de génération de ces termes d'indexation (Hollink *et al.*, 2004 ; McNamee & Mayfield, 2004). Dans certains cas, de meilleures performances peuvent être obtenues en combinant cette approche avec des méthodes basées sur la morphologie, soit directement à l'indexation en générant les *n*-grammes après traitement morphologique (Hollink *et al.*, 2004) ou à la recherche en fusionnant les résultats obtenus avec l'approche par *n*-grammes et l'approche morphologique (Savoy, 2003).

Face à ces différences de performances dues à la morphologie, il est difficile d'imaginer un système de RI générique qui puisse traiter uniformément toutes les langues avec leurs particularités et atteindre, en même temps, des performances optimales (*i.e.* précision moyenne maximale). Certes, le mot constitue l'unité lexicale la plus représentative pour la plupart des langues européennes, mais le degré du traitement morphologique nécessaire varie d'une langue à l'autre. Par ailleurs, si pour ces langues européennes, l'identification de ces mots ou unités lexicales est relativement aisée, cette tâche s'avère plus complexe pour certaines langues asiatiques comme le chinois ou le japonais. Dans ces deux langues, les mots ne sont pas explicitement marqués ; une phrase est une succession de symboles sans espaces (voir annexes A et B). L'indexation d'un document rédigé dans une langue présentant une telle caractéristique nécessite une étape préliminaire qui vise à segmenter le texte afin d'identifier les unités d'indexation nécessaires pour représenter le document. Cette segmentation, habituellement automatique, peut s'opérer comme une subdivision en blocs de taille fixe ou selon des informations lexicales en utilisant parfois des données statistiques. De plus, ces langues asiatiques se distinguent par d'autres aspects. Par exemple, le nombre

d'idéogrammes s'avère très élevé (par exemple plus de 13 000 pour le chinois traditionnel ou 7 700 pour le chinois simplifié)[11] comparé à nos 26 lettres. La langue japonaise combine à l'écriture chinoise (nommé *kanji*) trois autres systèmes, à savoir le *katakana* et le *hiragana* (deux systèmes phonétiques), et notre alphabet latin (utilisé pour indiquer certains nombres ou pour désigner des noms propres comme « Honda »). En coréen (voir annexe C), les mots sont explicitement délimités mais, à l'exemple de l'allemand, cette langue possède de très nombreux mots composés générés par concaténation et adjonction de divers mots simples ou racines et suffixes.

Étant donné ces différentes observations, nous souhaitons savoir dans cette thèse si les méthodes et les techniques développées pour les langues européennes peuvent être adaptées aux langues asiatiques. Dans ce but, nous avons choisi de travailler sur trois langues asiatiques, à savoir le chinois, le japonais et le coréen, ainsi que la langue anglaise. Ce choix est motivé par deux raisons principales : premièrement, ces trois langues asiatiques présentent des caractéristiques morphologiques et visuelles différentes de la langue anglaise et de la plupart des langues européennes. Elles représentent donc un réel défi pour la RI. Deuxièmement, les campagnes d'évaluation de NTCIR fournissent un environnement idéal pour mener nos recherches. En effet, ces conférences ont constitué de nombreuses collections-tests pour les langues chinoise, japonaise, coréenne et anglaise d'une part et d'autre part, ces campagnes offrent un terrain d'évaluation comparative qui permet de situer nos travaux par rapport aux autres systèmes de RI développés dans le domaine. Fort de ces deux raisons, notre premier objectif était de fournir des moteurs de recherche pour ces quatre langues avec la contrainte d'atteindre les meilleures performances possibles pour chacune d'elles (sections 2.1 et 2.2).

Au regard de la deuxième question, nous avons choisi de travailler dans le contexte biomédical. Comme nous l'avons démontré par quelques exemples dans la section précédente, ce domaine comporte de nombreux défis. Notre deuxième objectif consistait alors à évaluer et éventuellement à adapter les stratégies que nous avons développées pour la langue anglaise à NTCIR dans le but de proposer des solutions permettant de répondre aux particularités de ce domaine (section 2.3).

Enfin, concernant la dernière question, notre but consistera à jeter un regard critique sur quelques facettes liées à l'évaluation des systèmes de RI. Nous aborderons également les difficultés et les limites possibles que pourraient rencontrer nos stratégies de recherche face à certaines requêtes. Pour réaliser cette partie du travail, nous nous sommes appuyés sur une collection-test en langue française (section 2.4).

---

[11] Le chinois simplifié est utilisé en Chine continentale et à Singapour tandis que le chinois traditionnel est utilisé à Taiwan et Hong Kong.

1.3. *Organisation de cette thèse*

La suite de cette thèse est organisée de la façon suivante. Dans le reste de ce chapitre, nous allons présenter quelques concepts et définitions liés à la RI (section 1.4) suivi de la méthodologie d'évaluation adoptée dans nos différents travaux (section 1.5). Ensuite, le chapitre 2 sera consacré à la présentation des articles que nous avons publiés et qui forment l'essentiel de cette thèse. Enfin, le chapitre 3 présente la conclusion où nous exposerons nos principales contributions au domaine de la RI ainsi que les perspectives et travaux futurs.

1.4. *Concepts et définitions*

Nous présentons dans cette section les concepts et les définitions de certains aspects de la RI qui sont utilisés librement dans cette dissertation.

Recherche d'information (RI)

Une des premières définitions de la RI est donnée par Salton (1968). L'auteur la décrit comme étant le domaine qui concerne la structure, l'analyse, l'organisation, le stockage, la recherche et l'extraction de l'information. L'objectif principal de la RI est d'assister les utilisateurs à trouver, dans une collection volumineuse de documents, l'information qu'ils cherchent à partir de leurs besoins d'information généralement exprimé de façon vague et imprécise. Si initialement cette définition de la RI concernait uniquement le contenu textuel des documents ; avec l'ère du numérique, elle s'applique aujourd'hui à une large variété de supports d'information comme les images, les bandes sonores, les vidéos, les messages électroniques, les livres ou encore les documents multimédias et multilingues. Dans cette thèse, la référence à la RI concerne uniquement la recherche dans un contenu textuel.

Un système de RI classique comprend essentiellement trois processus : la représentation du contenu des documents, la représentation du besoin d'information de l'usager et enfin le processus d'appariement de ces deux représentations. La figure 1 illustre les différentes composantes d'un système de RI. Comme on peut le constater sur cette figure, un processus de rétroaction peut venir supporter la tâche de recherche en offrant à l'usager la possibilité de raffiner son besoin d'information en fonction des résultats que lui a présenté le système lors d'une recherche précédente. Ce dernier processus peut s'effectuer automatiquement comme le proposent Buckley *et al.* (1996) en exploitant directement les premiers documents retrouvés sous l'hypothèse qu'ils sont pertinents, il s'agit alors de pseudo-rétroaction.

Le fonctionnement d'un système de RI se déroule en deux phases. La première consiste à indexer la collection de documents pour laquelle on souhaite fournir un service de recherche. Cette étape, généralement réalisée hors ligne, peut s'effectuer selon trois méthodes : manuelle, automatique ou semi-automatique. Avec

l'indexation manuelle, chaque document est analysé par un spécialiste du domaine correspondant qui détermine, selon ses connaissances et souvent selon un vocabulaire contrôlé, les mots-clés qui lui semblent les plus significatifs pour représenter le contenu du document. A l'inverse, l'indexation automatique est complètement réalisée par la machine. Les techniques existantes pour représenter le contenu des documents vont d'une simple extraction de mots à des analyses linguistiques ou statistiques afin d'extraire les unités d'indexation les plus représentatives. Pour l'indexation semi-automatique, les deux approches précédentes sont combinées. Dans ce cas, le processus d'indexation tiendra compte non seulement du contenu du document mais aussi des descripteurs qui auraient été manuellement attribués par un indexeur humain. L'efficacité relative de ces trois approches d'indexation a fait l'objet de nombreuses investigations et constitue un des sujets que nous avons abordé dans cette thèse. A la fin de cette première phase, un index est constitué et en général stocké sous forme d'un fichier inversé qui, à partir de mots clés, permet de retrouver rapidement des documents.

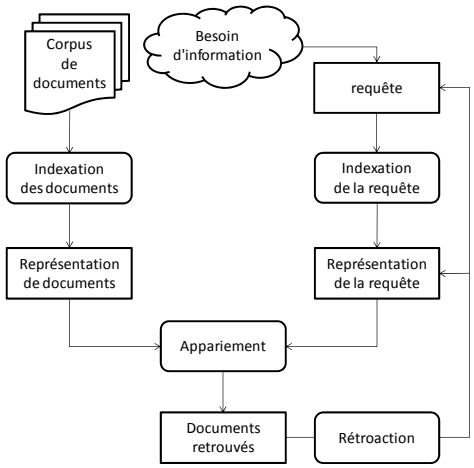

Figure 1. Schéma classique du processus de RI

La seconde phase du fonctionnement d'un système de RI est interactive et se charge de recevoir et de traiter les besoins d'informations des utilisateurs. Ainsi, pour une requête donnée, le processus d'indexation, souvent identique à celui des documents, analysera ce besoin d'information afin d'extraire une représentation compatible à celles des documents. Ensuite le processus d'appariement, basé sur cette représentation de la requête, va dépister et classer par ordre de pertinence tous les documents correspondant à cette requête. Enfin, le résultat de ce processus est présenté à l'utilisateur sous forme d'une liste de documents triés par ordre décroissant de pertinence.

Pertinence

La pertinence détermine le degré d'adéquation d'un document par rapport au besoin d'information de l'usager. De nombreux systèmes de RI tentent de modéliser cette notion de pertinence ainsi que l'incertitude sous-jacente afin de répondre au mieux aux requêtes de l'utilisateur. Cependant, la pertinence est une notion subjective puisqu'un document jugé pertinent par une personne ne le sera pas forcément par une autre ou par la même personne à un autre moment (Saracevic, 1975). Autrement dit, cette pertinence dépend fortement de l'interprétation que font les usagers de l'information contenue dans le document mais elle dépend aussi de leurs préférences, de leurs connaissances et a priori du contexte. Toutefois, l'étendue de cette subjectivité demeure relativement faible pour invalider les évaluations basées sur les jugements de pertinence puisque ceux-ci sont généralement réalisés par des experts (van Rijsbergen, 1979). De plus, Voorhees (1998) a démontré que l'utilisation d'un ensemble de jugements de pertinence différent affecte la performance absolue mais pas la performance relative. Autrement dit, un ensemble différent de jugements de pertinence peut conduire à des performances différentes, mais les systèmes ayant obtenu les meilleurs résultats avec un ensemble de jugements de pertinence tendent à fournir aussi de très bons résultats avec un autre ensemble de jugements de pertinence. Ainsi, le classement des systèmes ne change pas avec des ensembles différents de jugements de pertinence. Pour plus d'informations sur la notion de pertinence, l'article de Saracevic (2006) retrace l'évolution de ce concept au cours des trois dernières décennies.

Modèle de recherche

Un modèle de recherche est un formalisme du processus d'appariement entre un document et une requête. Il est au cœur du système de recherche et se traduit généralement par une formule mathématique permettant de mesurer le degré de ressemblance ou de similarité entre les documents et la requête. Il existe essentiellement quatre classes de modèles de RI, à savoir les modèles booléens, les modèles logiques, les modèles vectoriels et les modèles probabilistes. Van Rijsbergen (1979) et Ihadjadene (2004) proposent une description détaillée de chacune de ces classes. Dans les paragraphes suivants, nous présentons une brève description du fondement des modèles que nous avons utilisés dans nos évaluations.

Modèle vectoriel

Le modèle vectoriel, proposé à la fin des années 1960 par Salton et ses collaborateurs (Salton, 1971 ; Salton & McGill, 1983) dans le système SMART, est l'un des plus populaires dans le domaine de la RI. Le principe de ce modèle s'appuie sur une intuition géométrique du processus d'appariement entre un document et une requête. Dans cette approche, ces derniers sont représentés comme des vecteurs appartenant à un espace vectoriel multidimensionnel. Chaque terme d'indexation est considéré comme une dimension de cet espace. La pertinence d'un document par rapport à une requête est relative aux positions respectives de ces

vecteurs dans cet espace. La figure 2 présente l'exemple d'un espace tridimensionnel comprenant un vecteur document (D) et un vecteur requête (Q). Les trois dimensions sont associées aux termes « drogue », « dure » et « douce ». Selon le positionnement du vecteur document, le terme « douce » possède un poids plus important que le terme « dure » tandis que le vecteur requête (Q) concerne les termes « drogue » et « dure » à poids égal. Cette représentation géométrique permet d'expliquer plus facilement le fonctionnement du modèle pour des non spécialistes.

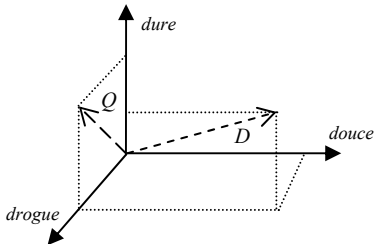

Figure 2. Représentation du modèle vectoriel

Le classement des documents retrouvés en réponse à une requête est déterminé par une mesure de similarité (ou plus généralement le score) entre le document et la requête. Cette similarité peut être calculée suivant le produit interne selon l'équation 1 ou par exemple selon le cosinus de l'angle séparant le document et la requête suivant l'équation 2. Dans ce cas, cette mesure de similarité est nulle si les vecteurs sont orthogonaux, et est égale à un si l'angle est nul. D'autres coefficients ont été proposés pour mesurer cette similarité comme les coefficients de *Dice* ou de *Jaccard* (van Rijsbergen, 1979).

$$score(D,Q) = sim\left(\overrightarrow{D_j},\vec{Q}\right) = \overrightarrow{D_j}\cdot\vec{Q} = \sum_{i=1}^{|Q|} w_{ij}\cdot w_{iq} \qquad (1)$$

$$score\left(\overrightarrow{D_j},\vec{Q}\right) = \frac{\overrightarrow{D_j}\cdot\vec{Q}}{\left\|\vec{D_j}\right\|\times\left\|\vec{Q}\right\|} = \frac{\sum_{i=1}^{|Q|} w_{ij}\cdot w_{iq}}{\sqrt{\sum_{i=1}^{|Q|} w_{ij}^2}\times\sqrt{\sum_{i=1}^{|Q|} w_{iq}^2}} \qquad (2)$$

La pondération des termes dans le document ($w_{i,j}$) ou la requête ($w_{i,q}$) tient compte généralement de trois critères : 1) le nombre d'occurrences du terme dans le document ou la requête ; 2) l'importance du terme dans la collection mesurée par sa fréquence documentaire, c'est-à-dire par le nombre de documents indexés par ce terme ; 3) enfin, un facteur de normalisation qui permet de prendre en compte la longueur des documents.

De nombreuses variantes du modèle vectoriel ont été mises au point, dont le modèle « Lnu », qui vise à mieux discriminer les documents selon leur longueur (Singhal *et al.*, 1996). Cette stratégie s'est avérée particulièrement efficace pour la langue chinoise (Chung *et al.*, 2006) et pour d'autres langues européennes et asiatiques (Savoy, 2005).

Finalement, bien que l'approche vectorielle repose sur une base théorique relativement faible par rapport aux modèles probabilistes (voir ci-dessous), elle constitue néanmoins le moteur de plusieurs systèmes de recherche commerciaux comme ceux basés sur LUCENE (Hatcher & Gospodnetić, 2005). Dans le milieu académique, le modèle vectoriel classique ($tf \cdot idf$) est souvent utilisé comme base de comparaison pour évaluer les nouvelles stratégies de recherche.

Modèle probabiliste (Okapi BM25)

Contrairement aux modèles vectoriels, les approches probabilistes possèdent un fondement théorique rigoureux. Le modèle probabiliste proposé par Robertson (1977) s'appuie sur le principe du classement probabiliste (*Probability Ranking Principle*) qui énonce qu'un système retournant les documents dans l'ordre décroissant de leur probabilité de pertinence par rapport à la requête opère de manière optimale. Une définition précise et plus formelle de ce principe est également donnée dans par Sparck-Jones *et al.* (2000a). L'idée fondamentale de cette approche est de classer les documents retrouvés par probabilité de pertinence. Plus précisément, il s'agit de déterminer la probabilité de retrouver un document D_j, et que celui-ci appartient à l'ensemble des documents pertinents (R) ou non-pertinents (nR). En notant ces probabilités respectivement $P(R|D_j)$ et $P(nR|D_j)$, la similarité entre le document et la requête est définie comme étant le rapport de ces deux probabilités, soit :

$$score(D,Q) = \frac{P(R|D_j)}{P(nR|D_j)} \tag{3}$$

mais comme ces deux probabilités ne peuvent être calculées directement, le théorème de Bayes[12] est appliqué pour obtenir :

$$score(D,Q) = \frac{P(D_j|R) \cdot P(R)}{P(D_j|nR) \cdot P(nR)} \tag{4}$$

dans laquelle $P(D_j|R)$ est la probabilité de sélectionner le document D parmi l'ensemble des documents pertinents (R). De plus, $P(R)$ est la probabilité que le document sélectionné au hasard dans le corpus soit pertinent. Les significations

[12] Selon le théorème de Bayes, $P(A|B) = (P(B|A) \cdot P(A))/P(B)$

liées à $P(D_j|nR)$ et $P(nR)$ sont analogues et complémentaires. Comme les valeurs de $P(R)$ et $P(nR)$ sont les mêmes pour tous les documents de la collection, ils ne changent donc pas le classement final. On peut alors écrire :

$$score(D_j, Q) \approx \frac{P(D_j|R)}{P(D_j|nR)} \quad (5)$$

Les différentes approches pour estimer $P(D_j|R)$ et $P(D_j|nR)$ ont conduit au développement de multiples modèles. Parmi ces modèles basés sur la distribution de Poisson pour estimer les probabilités sous-jacentes, Robertson & Walker (1994) ont expérimenté, en utilisant le système Okapi développé à *City University of London*, diverses stratégies de pondération tenant compte de la fréquence du terme ainsi que de la longueur des documents. La formule BM25 (Sparck-Jones *et al.*, 2000b), aujourd'hui communément appelé Okapi, constitue l'aboutissement de cette série d'expériences. Une formulation simplifiée de cette pondération est reprise dans les publications annexées à cette dissertation.

Modèle de langue

Dans les approches probabilistes comme celle décrite ci-dessus, la pertinence (R) d'un document (D) par rapport à un besoin d'information de l'usager (Q) est généralement modélisée par une probabilité de pertinence, notée $P(R|Q, D)$. Celle-ci est en principe calculée selon l'hypothèse que la distribution des termes dans les documents pertinents suit une loi de probabilité donnée, comme la loi de Poisson. Or, dans les modèles de langue, cette pertinence n'est pas directement modélisée mais elle est interprétée par la probabilité que la requête puisse être générée par le modèle de langue du document. Les probabilités sont calculées directement à partir du modèle de langue statistique du document et ne dépendent a priori d'aucune distribution. Dans cette approche, le classement des documents pour une requête donnée est déterminé par la probabilité du modèle de langue de chaque document dans la collection à engendrer cette requête. Ainsi, le score d'un document (D_j) face à une requête (Q) est défini par la probabilité que cette requête puisse être générée par le modèle de langue de ce document (M_{D_j}), soit :

$$score(D_j, Q) = P(Q|M_{D_j}) \quad (6)$$

En considérant la requête (Q) comme une suite de termes $(t_1 t_2 \cdots \mathrm{t_n})$ indépendants[13], on peut alors écrire :

[13] L'indépendance des termes est une hypothèse simplificatrice très répandue dans le développement de modèles de RI. Dans les modèles de langue, cette hypothèse se traduit par l'utilisation d'un modèle unigramme qui constitue le modèle standard des modèles de langue en RI. Les modèles de langue d'ordre supérieur (bigramme ou trigramme) ont été proposés la première fois en RI par Song & Croft (1999) et Miller *et al.* (1999).

$$score(D_j, Q) = P\left(t_1 t_2 \cdots t_n \middle| M_{D_j}\right) = \prod_{i=1}^{n} P\left(t_i \middle| M_{D_j}\right) \quad (7)$$

$$P\left(t_i | M_{D_j}\right) = \frac{tf(t_i, D_j)}{|D_j|} \quad (8)$$

où l'estimation de vraisemblance maximale (ou la fréquence relative, calculée selon l'équation 8 comme étant le rapport entre la fréquence absolue du terme dans le document et la longueur de celui-ci en nombre de termes d'indexation distincts) constitue le moyen le plus simple pour estimer les probabilités $P(t_i|M_{D_j})$. Cependant, un terme d'une requête absent dans un document engendrera une probabilité nulle et le document ne sera donc pas dépisté. Afin d'éviter ce problème, plusieurs méthodes de lissage ont été proposées pour attribuer des probabilités non nulles à des termes n'apparaissant pas dans les documents (Zhai & Lafferty, 2004 ; Hiemstra, 2001). Par exemple, Hiemstra (1998) propose de recourir à une interpolation linéaire combinant le modèle de langue du document avec le modèle de langue de la collection, soit :

$$P\left(t_i | M_{D_j}\right) = (1 - \lambda) P_{ML}(t_i | M_C) + \lambda P_{ML}(t_i | M_{D_j}) \quad (9)$$

où λ est un paramètre de lissage, et les probabilités $P_{ML}(t_i|M_{D_j})$ et $P_{ML}(t_i|M_C)$ sont des estimations de vraisemblance maximale selon respectivement le document et la collection. Les formules exactes de ce modèle sont données dans les publications annexées à cette thèse.

Enfin, le paradigme modèle de langue, issu du domaine de la reconnaissance de la parole, a été premièrement introduit en RI par Ponte et Croft (1998) où une performance d'environ 20 % supérieure au modèle vectoriel ($tf \cdot idf$) a été observée. Actuellement, il est admis que les modèles de langue offrent une performance comparable, voire supérieure au modèle Okapi.

Divergence from Randomness (DFR)

Amati & van Rijsbergen (2002) ont proposé un paradigme permettant d'engendrer de nouveaux modèles probabilistes pour la RI. Ce paradigme, nommé *Divergence from Randomness* (DFR), est basé sur une mesure de gain associé au dépistage d'un document contenant un terme de la requête. Dans ce cas, le score d'un document par rapport à une requête est donnée par :

$$score(D_j, Q) = \sum_{i=1}^{|Q|} w_{ij} \cdot w_{iq} \quad (10)$$

où $w_{i,q}$ correspond généralement à la fréquence du terme dans la requête et $w_{i,j}$ est le poids du terme dans le document. Ce poids combine deux mesures d'informations suivant la formule :

$$w_{ij} = Inf_{ij}^{1} \cdot Inf_{ij}^{2} \tag{11}$$

Le principe fondamental de cette approche est le suivant : plus la probabilité d'occurrence d'un terme dans un document diverge de sa probabilité d'occurrence dans la collection, plus le degré d'informativité de ce terme est élevé pour ce document. Ainsi, le poids du terme est inversement lié à la probabilité de la fréquence de ce terme dans le document. Ce poids est obtenu par la modélisation des occurrences suivant une distribution aléatoire, par exemple une loi binomiale, une loi de Poisson, l'inverse de la fréquence documentaire ou la statistique de Bose-Einstein. Une faible probabilité correspondra donc aux termes spécifiques pour lesquels l'informativité mesurée par la formule 12 sera grande. A l'inverse, des termes possédant une forte probabilité d'occurrence, selon la distribution aléatoire choisie, sont des termes généraux pour lesquels cette informativité sera faible.

$$Inf_{ij_1}^{1} = -log_2 Prob_{ij}^{1} \tag{12}$$

$$Inf_{ij_1}^{2} = 1 - Prob_{ij}^{2} \tag{13}$$

D'autre part, Amati & van Rijsbergen (2002) ont inclus une composante de risque qui vise à normaliser la mesure d'informativité estimée par la formule 12. Ce facteur de normalisation, calculé suivant l'équation 13, mesure le risque d'accepter un terme comme un bon descripteur du document. Dans cette dernière équation, la probabilité ($Prob_{\mathrm{i,j}}^{2}$) est calculée en observant l'ensemble des documents indexés par le terme. Elle indique la probabilité de rencontrer une nouvelle occurrence du terme dans un document donné sachant que l'on a déjà dénombré tf occurrences. Cette probabilité est modélisée par la loi des successions de Laplace ou le ratio de deux processus de Bernoulli (se référer aux articles de cette thèse pour les formules détaillées).

$$tfn_{ij} = tf_{ij} \cdot log_2 \left[1 + \frac{c \cdot mean\ dl}{l_i}\right] \tag{14}$$

Enfin, le calcul du poids du terme dans le document selon la formule 11 est précédé par la normalisation de la fréquence de ce terme suivant l'équation 14 afin de tenir compte du fait que les documents ne possèdent pas tous la même longueur. Dans cette formule, tf_{ij} correspond à la fréquence originale du terme dans le document, l_i à la longueur du document, $mean\ dl$ à la longueur moyenne des

documents du corpus et c à un paramètre permettant de contrôler la longueur moyenne des documents. Ce paramètre est généralement fixé empiriquement et dépend de la collection utilisée. He & Ounis (2005) ont montré que les résultats peuvent être améliorés avec l'ajustement de ce paramètre. Ces auteurs ont également proposé une approche automatique et indépendante de la collection pour fixer ce paramètre. Plus récemment, Amati (2006) a proposé de nouveaux modèles basés sur ce paradigme, qui ne dépendent pas de paramètres particuliers.

1.5. *Méthodologie d'évaluation*

Le domaine de la RI repose sur une ancienne tradition empirique qui remonte à la fin des années soixante avec le projet *Cranfield* (Cleverdon, 1967). Ces travaux ont eu une influence considérable sur la méthodologie de l'évaluation des systèmes de RI et le « paradigme de *Cranfield* » est devenu sans doute le résultat le plus important de ce projet. En effet, les expériences de Cleverdon (1967) sur les techniques d'indexation (essentiellement sur les diverses variantes de l'indexation manuelle) ont permis d'établir un protocole d'évaluation constituant aujourd'hui, avec quelques adaptations, la base des campagnes d'évaluation de TREC, NTCIR et CLEF. En plus de quelques hypothèses simplificatrices (Voorhees, 2002)[14], l'essentiel de ce protocole comprend deux composantes, à savoir la collection-test et les mesures de performance.

Collection-test

Une collection-test correspond à un environnement d'expérimentation contrôlé et composé d'un corpus de documents, d'un ensemble de requêtes décrivant les besoins d'informations et d'un ensemble de jugements de pertinence indiquant pour chaque requête les documents jugés pertinents et non pertinents. L'ensemble des requêtes ainsi que les jugements de pertinence relatifs sont établis par des usagers réels et non par une machine (artificiellement). Les expériences de *Cranfield* ont permis de construire la première collection-test en RI composé de 1400 notices bibliographiques dans le domaine de l'aéronautique, 225 requêtes et un ensemble exhaustif de jugements de pertinence. Pour des collections volumineuses, l'établissement d'un ensemble de jugements de pertinence complet (une des hypothèses de *Cranfield*) n'est pas possible pour des raisons de temps et de coûts. Pour pallier ce problème dans les campagnes d'évaluation, on utilise la méthode du « *pooling* ». Ce procédé consiste à regrouper un nombre limité (par exemple une centaine) des meilleurs documents retrouvés par chaque système et pour chaque requête. Après avoir éliminé les éventuels doublons, ce sous-ensemble est ensuite jugé par des experts ou les personnes ayant soumis la requête. Les documents qui

[14] La pertinence d'un document peut être approximée par une mesure de similarité entre le document et la requête ; la valeur de pertinence d'un document est indépendante du nombre de documents pertinents ; la connaissance exhaustive des documents pertinents.

n'ont pas été jugés seront considérés non pertinents. Cette procédure contrevient à l'hypothèse de *Cranfield* mais Voorhees (2002) a démontré qu'avec des systèmes diversifiés cette méthode du « *pooling* » représente une bonne approximation et permet de produire des jugements non biaisés, nécessaires à toute évaluation *comparative* de systèmes. La performance ne peut être mesurée de façon absolue mais toujours en comparaison d'autres approches.

Mesures de performance

Parmi les critères d'évaluation d'un système de RI présentés dans le paradigme de *Cranfield* (Cleverdon, 1967), la précision et le rappel constituent actuellement les mesures de référence pour évaluer l'efficacité des systèmes de RI. Ces deux mesures combinées permettent d'évaluer la capacité d'un système à retourner les bonnes réponses en excluant en même temps les réponses non pertinentes. En particulier pour une requête donnée, la précision représente la proportion du nombre de document pertinents parmi l'ensemble des documents retournés, tandis que le rappel représente la proportion du nombre de documents pertinents retrouvés parmi l'ensemble des documents pertinents. Les travaux de *Cranfield* ont montré que ces deux mesures sont inversement liées (Gordon & Kochen, 1989). Ainsi, lorsque la précision augmente le rappel a tendance à diminuer. Ces deux mesures possèdent l'avantage de répondre à des besoins différents. Alors qu'un système sur le Web tend à maximiser la précision afin de fournir uniquement quelques bonnes réponses au début de la liste retournée par la machine, un système de recherche dans un contexte spécifique, comme le domaine légal, s'intéressera plus à optimiser le niveau de rappel pour retrouver toutes les décisions des tribunaux limitant ou renversant une cause précise.

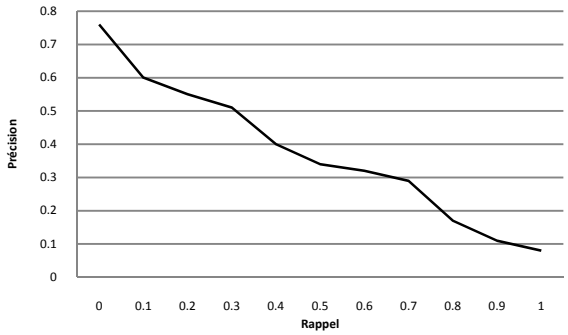

Figure 3. Exemple de courbe précision-rappel.

Pour évaluer une liste ordonnée de résultats pour une requête, on calcule la précision à différents niveaux de rappel. Ainsi, on obtient des paires de valeurs qui permettent d'établir une courbe de précision-rappel (figure 3). Certes, le graphe

obtenu peut être informatif mais on désire parfois résumer cette illustration ou ensemble de mesures avec une seule valeur. A cet effet, plusieurs solutions ont été proposées dont la précision moyenne (ou MAP) que nous aborderons plus en détail avec d'autres aspects liés à l'évaluation dans l'article relatif à la section 2.4.

Les travaux que nous avons menés dans cette thèse se situent dans le cadre des campagnes d'évaluation de NTCIR, TREC et CLEF. Les collections-tests constituées lors de ces conférences ainsi que la méthodologie adoptée pour la comparaison et l'évaluation des stratégies de recherche nous ont permis de réaliser nos différentes expérimentations.

2. Présentation des articles

Le contenu de cette dissertation est basé sur les quatre publications suivantes :

1. Samir Abdou, Jacques Savoy, « Report on CLIR Task for the NTCIR-5 Evaluation Campaign », In *Proceedings of the Fifth NTCIR Workshop Meeting*, Tokyo, 2005, pages 44-51.
2. Samir Abdou, Jacques Savoy, « Statistical and Comparative Study of Various Indexing and Search Models », In *Proceedings of the Third Asia Information Retrievals Symposium (AIRS 2006)*, Singapore, 2006, pages 362-373.
3. Samir Abdou, Jacques Savoy, « Searching in MEDLINE: Query Expansion and Manual Indexing Evaluation », *Information Processing and Management* (*to appear*).
4. Samir Abdou, Jacques Savoy, « Considérations sur l'évaluation de la robustesse en recherche d'information », In *ACTES de la quatrième Conférence de Recherche d'Information et Applications (CORIA 2007)*, Saint-Étienne, France, 2007, pages 5-20.

Ces publications abordent diverses questions au regard des motivations et objectifs de cette thèse. Dans les sections suivantes, nous allons donner une description de chacune de ces contributions en exposant d'abord les interrogations qu'elles soulèvent et éventuellement l'état des connaissances. Nous présenterons ensuite l'approche adoptée pour apporter des réponses. Enfin, nous terminerons cette description en présentant les principaux résultats de la contribution.

2.1. *Report on CLIR Task for the NTCIR-5 Evaluation Campaign*

Cet article présente les travaux menés dans le cadre de notre participation à la cinquième campagne d'évaluation de NTCIR (Kishida *et al.*, 2005). Nous avons décrit dans cette publication nos expériences sur la RI monolingue, bilingue et

multilingue pour les langues chinoise, japonaise, coréenne et anglaise. Toutefois, nos efforts se sont essentiellement focalisés sur la partie monolingue car cela constituait une étape fondamentale pour proposer des systèmes de RI bilingues (où les requêtes sont données en langue anglaise et les documents dans une des trois autres langues asiatiques) et multilingues (où la recherche s'effectue à la fois dans les quatre langues à partir de requêtes formulées en langue anglaise).

Lorsqu'on aborde le développement ou l'adaptation d'un système de RI pour une nouvelle langue, nous devons répondre d'abord à deux questions essentielles. D'une part, quelles sont les unités d'indexation (mots, lemmes, morphèmes, *n*-grammes, etc.) pouvant le mieux représenter le contenu des documents (et requêtes) de cette langue et d'autre part, quels modèles de RI adopter. Ces deux questions étant abordées plus en détails dans la prochaine section, nous n'en donnerons ici que les grandes lignes. Pour la première question, la réponse s'avère relativement aisée pour la langue anglaise ; les mots sont clairement délimités. En revanche, cette tâche est un peu plus complexe pour les trois langues asiatiques. Ainsi pour les langues chinoise et japonaise, nous avons comparé deux stratégies d'indexation, à savoir l'indexation par bigrammes et l'indexation par mots. Afin d'identifier les mots pour ces deux langues, nous avons utilisé l'analyseur morphologique Chasen (Matsumoto *et al.*, 1999) pour la langue japonaise et l'outil de segmentation MTSeg[15] pour la langue chinoise. Pour cette dernière langue, nous avons également construit un index combinant les représentations par unigrammes et par bigrammes. Pour la langue coréenne, seule l'approche par bigrammes a été considérée. Outre l'objectif de comparer ces différentes stratégies d'indexation, nous nous sommes intéressés à évaluer l'effet de la fusion des listes de résultats obtenus avec ces différentes représentations ; nous aborderons ce point plus loin. Pour la seconde question, nous avons évalué onze modèles de recherche. En plus du modèle Okapi (Robertson *et al.*, 2000) et de neuf modèles vectoriels, nous avons implémenté deux modèles DFR (Amati & van Rijsbergen, 2002), à savoir le modèle PB2 que nous avons utilisé pour les trois langues asiatiques et le modèle I(*n*)L2 utilisé pour la langue anglaise. Ces choix ont été motivés par nos expériences préliminaires réalisées sur les collections-tests de NTCIR-4.

En apportant des réponses à ces deux interrogations, nous disposons ainsi d'un système de base pour chaque langue. Dans les campagnes d'évaluation de CLEF, TREC et NTCIR, de nombreux participants recourent à l'utilisation d'outils supplémentaires afin d'augmenter la performance de leurs systèmes de base. Par exemple, la pseudo-rétroaction (ou expansion de requêtes) est une stratégie largement utilisée dans ces différentes campagnes. Ainsi, nous avons adopté cette technique dans le but d'améliorer nos moteurs de recherche.

A cet effet, nous avons proposé et évalué une nouvelle stratégie pour l'expansion de requêtes. Cette méthode, que nous avons baptisé « *idf query expansion* » (ou IDFQE), est basée sur l'inverse de la fréquence documentaire (idf) pour

[15] Disponible sur http://www.mandarintools.com

sélectionner et pondérer les termes importants. La mesure idf a été proposée initialement par Sparck-Jones (1972) comme moyen de mesurer la spécificité d'un terme ; elle repose sur le nombre de documents indexés par le terme en question. Plusieurs modèles de recherche utilisent cette mesure comme facteur de pondération global. Une des formes les plus citées dans la littérature en RI, et que nous utilisons dans notre approche, est calculée comme suit :

$$idf_j = log\left(\frac{N}{df_j}\right) \tag{15}$$

avec N correspondant au nombre de documents de la collection et df_j à la fréquence documentaire du terme j (le nombre de documents indexés par le terme). Ainsi, un terme apparaissant dans peu de documents aura une forte valeur idf tandis qu'un terme indexé par beaucoup de documents possèdera une faible valeur idf. Dans le premier cas, il s'agit essentiellement de termes spécifiques comme les noms propres de personnes, de villes ou d'entreprises mais aussi, quelquefois, de termes généraux mal orthographiés. Dans le deuxième cas, les termes véhiculent peu de sémantique et correspondent à des termes très fréquents ou des outils grammaticaux qui n'ont pu être filtrés lors du processus d'indexation par une liste de mots-outils. Une analyse détaillée de l'idf est donnée par Robertson (2004) et une revue d'histoire ainsi que l'impact qu'a eu cette mesure sur la RI est présentée par Harman (2005b).

Étant donné ces propriétés et caractéristiques de l'idf, il serait judicieux d'utiliser cette mesure afin de reformuler les requêtes initiales en sélectionnant et en pondérant les termes spécifiques qui apparaissent dans les premiers documents retournés par le système. De même que la pseudo-rétroaction (Buckley *et al.*, 1996), notre approche repose sur l'hypothèse que les premiers documents de la liste retournée sont pertinents. La description précise de l'algorithme, dont nous venons d'esquisser le fondement, est donnée dans l'article (voir aussi la section 2.3 et l'article correspondant). Une comparaison de cette stratégie avec celle de Rocchio (Buckley *et al.*, 1996) est également fournie.

Une autre approche pouvant améliorer la performance est la fusion de listes de résultats. L'idée de cette stratégie est basée sur l'observation suivante : certains systèmes ayant des performances globales similaires retrouvent des documents pertinents différents. Dans ce cas, ces systèmes peuvent être complémentaires et la fusion de leurs résultats pourrait apporter une amélioration de la qualité de recherche. Dans cette optique, pour chaque langue nous avons évalué différentes stratégies de fusion en exploitant les listes de résultats intermédiaires, à savoir celles obtenues après expansion de requêtes avec les modèles Okapi, PB2 et I(*n*)L2, en considérant les différentes stratégies d'indexation.

Ces premières expériences permettent de constater que l'indexation par mots s'avère légèrement supérieure qu'une approche par bigrammes de lettres pour les langues chinoise et japonaise. Par ailleurs, l'expansion de requêtes s'est révélée très

efficace pour améliorer les performances, en particulier notre nouvelle approche apporte une performance supérieure dans 30 cas sur 36 (voir les tables 4 et 5 dans l'article) que l'approche de Rocchio (Buckley *et al.*, 1996). De plus, les améliorations étaient très souvent statistiquement significatives.

Enfin, par rapport aux systèmes de recherche des autres participants à cette campagne d'évaluation, les résultats obtenus avec nos moteurs de recherche étaient très encourageants avec le premier rang pour les langues japonaise et anglaise, le deuxième rang pour le chinois et le troisième rang pour le coréen (Kishida *et al.*, 2005). Plus récemment, les travaux de notre participation à la sixième campagne d'évaluation de NTCIR (Abdou & Savoy, 2007) confirment ces bons résultats avec des performances supérieures à la médiane pour toutes les requêtes, et ce pour les trois langues asiatiques auxquelles nous avons participé (Kishida *et al.*, 2007).

2.2. *Statistical and Comparative Study of Various Indexing and Search Models*

Les travaux présentés dans cet article font suite à nos diverses investigations menées lors de la cinquième campagne d'évaluation de NTCIR (Abdou & Savoy, 2005). Plus précisément, nous nous sommes intéressés à identifier la meilleure approche d'indexation pour représenter les documents et les requêtes pour les langues chinoise, japonaise et coréenne. En plus de connaitre les stratégies de recherche les plus efficaces, deux raisons importantes nous ont encouragé à entreprendre cette étude. D'abord, peu de travaux comparant les différentes stratégies d'indexation pour ces trois langues ont été publiés. Ensuite, la plupart des études antérieures évaluent les différentes approches possibles à l'aide d'un nombre très limité de modèles de recherche. Enfin, étant donné le nombre de paramètres et de traitements sous-jacents possibles (requêtes structurées, expansion de requêtes, combinaisons diverses), la comparaison directe ne permet souvent pas de connaître l'influence de ces diverses composantes.

Afin de représenter des documents rédigés en langue chinoise, japonaise ou coréenne, on distingue généralement trois stratégies d'indexation différentes. Premièrement, les méthodes peuvent se baser sur les caractères en découpant le texte en *n*-grammes. Ce choix présente l'avantage de ne pas requérir de connaissances linguistiques forcément différentes pour chaque langue naturelle d'une part et d'autre part, sa mise en œuvre s'avère fort simple. Ce type d'approche présente une qualité de réponse intéressante pour certaines langues européennes (Savoy, 2003 ; McNamee & Mayfield, 2003) mais nécessite un temps de réponse plus long (environ 10 fois plus important). Deuxièmement, on peut recourir à des connaissances linguistiques (dictionnaires ou lexiques, règles morphologiques ou syntaxiques) ou s'appuyer sur une étude statistique des corpus pour reconnaître et extraire les mots d'un texte. Cependant, la couverture limitée des dictionnaires et autres listes de mots engendre souvent des erreurs de segmentation. En effet, en présence de noms propres ou de mots plus spécifiques, la segmentation automatique en mots d'une phrase est sujette à erreurs. Une méthode permettant de résoudre ce problème nous

conduit enfin à une troisième approche, dite hybride, consistant à combiner les deux premières approches selon divers modes opératoires.

L'effet de ces différentes stratégies d'indexation sur la performance des systèmes de recherche a fait l'objet de quelques études, sans qu'un consensus clair et admis par tous soit trouvé. Ainsi pour la langue chinoise, Kwok (1999) indique qu'une indexation par bigrammes s'avère meilleure qu'une approche unigrammes (collections-tests TREC-5 et TREC-6) tandis que la combinaison des deux méthodes améliore la performance. Luk & Kwok (2002), s'appuyant sur plusieurs collections-tests (TREC-5, trec-6, NTCIR-2 et TREC-9), sont parvenus au même résultat avec un modèle de recherche différent. Ces auteurs ont également démontré la supériorité de l'indexation par bigrammes par rapport à une indexation par mots (la collection-test de TREC-6 étant dans ce cas une exception à cette règle). En revanche, en comparant ces deux approches, Nie & Ren (1999) ont obtenu des performances comparables (modèle vectoriel), la meilleure précision étant obtenue par une combinaison "unigrammes & bigrammes". Finalement, Nie *et al.* (2000) ont comparé cette approche combinée avec une représentation par mots. Dans ce cas, les performances obtenues se sont avérées similaires tandis que la combinaison unigrammes avec mots (longs ?) apporte une meilleure précision moyenne.

Pour la langue japonaise, Chen & Gey (2003) ont comparé la combinaison unigrammes et bigrammes avec une approche basée sur une segmentation automatique (ou indexation par mots). Basé sur la collection-test de NTCIR-3, ces auteurs ont obtenu une différence marginale entre ces deux approches (0,2802 contre 0,2758, soit 1,6 % de différence relative). Avec la collection-test NTCIR-2 et un modèle de langue, l'approche bigrammes s'est avérée similaire à une approche par trigrammes (McNamee, 2001). Sur la collection-test de NTCIR-4, Tomlinson (2004) a obtenu des performances similaires en comparant la représentation par mots à celle basée sur des *n*-grammes.

Pour la langue coréenne, Lee & Ahn (1996) ont proposé une approche par bigrammes obtenus à partir des mots mais après un traitement morphologique. Cette représentation s'est avérée supérieure à une approche par mots (sans aucun prétraitement). Par contre, cette forme d'indexation présente une performance comparable à une approche par morphèmes (décomposition linguistique des mots après traitement morphologique). A l'aide de la collection-test de NTCIR-4, Kwok *et al.* (2004) ont évalué deux approches, à savoir la représentation par des bigrammes d'une part et, d'autre part, l'indexation par mots après traitement morphologique. Ces auteurs ont montré que la première approche permet une meilleure précision moyenne que la seconde. Les travaux de Tomlinson (2004) tendent à confirmer cette constatation : la représentation par des bigrammes tend à offrir une meilleure performance qu'une indexation par mots décomposés.

Dans cet article, afin d'apporter une réponse claire et précise à notre problématique, nous avons comparé quatre stratégies d'indexation pour le chinois et le japonais (unigrammes, bigrammes, uni- et bigrammes et, finalement, les mots) et

trois pour le coréen (mots, bigrammes et morphèmes). Afin de consolider nos conclusions, nous avons effectué cette comparaison à l'aide de neuf modèles de recherche parmi les plus efficaces, soit six modèles vectoriels, un modèle de langue (Hiemstra, 2001), le modèle PB2 de DFR (Amati & van Rijsbergen, 2002) et le modèle Okapi (Robertson *et al.*, 2000). Nous avons complété nos travaux par une analyse statistique approfondie où nous avons comparé les résultats de quatre tests (test du signe, test de Student, test de Wilcoxon et le test basé sur le ré-échantillonnage aléatoire) couramment utilisés en RI. Enfin, nous avons limité nos expériences à l'utilisation de requêtes courtes construites à partir de la section titre des besoins d'information (*Topics*).

Basé sur ces neuf modèles de recherche et les corpus de la campagne d'évaluation NTCIR-5, nous avons évalué la précision moyenne obtenue par diverses formes d'indexation pour les langues chinoise, japonaise et coréenne. Les principaux résultats que nous pouvons tirer de ces expériences sont les suivants.

Le modèle probabiliste PB2 issu de la famille DFR offre souvent la précision moyenne la plus élevée, indépendamment de la langue ou de la forme d'indexation (*n*-grammes ou mots). Comme alternative, nous pouvons suggérer le modèle Okapi ou l'approche vectorielle « Lnu-ltc » (Singhal *et al.*, 1996). En effet, les différences de performance entre ces systèmes de dépistage n'étaient généralement pas statistiquement significatives, indiquant que la modification de quelques requêtes ou la modification du corpus (par exemple de dépêches d'agences à un ensemble de document juridiques) peut modifier les classements présentés.

Pour la langue chinoise, la meilleure stratégie d'indexation semble être une approche combinée « unigrammes & bigrammes ». Pour la langue coréenne, une indexation par bigrammes apporte la meilleure qualité de recherche. Par contre pour la langue japonaise, les indexations par bigrammes, segmentation automatique (Chasen) ou combinée « unigrammes & bigrammes » apportent des performances statistiquement similaires.

L'analyse des différents tests statistiques indique que les résultats du test de Student et de celui basé sur le ré-échantillonnage aléatoire sont fortement corrélés. Cette corrélation linéaire demeure forte entre le test de Student et celui de Wilcoxon et s'avère un peu moins forte entre le test du signe et celui de Student. Nous avons démontré de manière empirique que les conclusions de ces quatre tests sont généralement très consistantes.

Finalement, ces résultats se révèlent concordants avec les travaux réalisés récemment lors de notre participation à NTCIR-6 (Abdou & Savoy, 2007) où les évaluations portaient sur quatre collections-tests différentes (NTCIR-3, NTCIR-4, NTCIR-5 et NTCIR-6) avec un moteur de recherche identique pour chaque langue, contrainte qui atteste de la stabilité de nos moteurs de recherche pour ces trois langues asiatiques.

2.3. *Searching in MEDLINE: Query Expansion and Manual Indexing Evaluation*

Dans cet article, nous avons abordé la RI dans un contexte spécifique, à savoir le domaine de la biomédecine. Dans la section introductive (section 1.1), nous avons démontré par de nombreux exemples les problèmes et difficultés auxquelles ce domaine est confronté. Dans le but de proposer des solutions pouvant améliorer la qualité de recherche dans ce contexte, nous avons conduit une série d'expériences sur la collection-test de MEDLINE (Hersh *et al.*, 2005). Pour mener ces travaux, nous avons adopté notre moteur de recherche développé pour la langue anglaise à NTCIR-5 (Abdou & Savoy, 2005). Dans ce cadre, cette contribution aborde trois aspects.

Tout d'abord, nous nous sommes intéressés à l'évaluation de divers modèles de recherche afin de déterminer la meilleure approche. A cet effet, nous avons comparé sept modèles vectoriels et trois modèles probabilistes, à savoir le modèle Okapi (Robertson *et al.*, 2000), le modèle I(*n*)B2 de DFR (Amati & van Rijsbergen, 2002) et un modèle de langue (Hiemstra, 2001).

Ensuite, nous avons porté une attention particulière à la technique d'indexation (manuelle *vs.* automatique). Pour situer le contexte, rappelons d'abord que chaque article indexé dans la base MEDLINE est soumis à un contrôle éditorial pour corriger les fautes d'orthographe d'une part et d'autre part, pour indexer manuellement l'article en lui attribuant des descripteurs MeSH[16] (*Medical Subject Headings*). L'objectif de ces opérations est d'augmenter la consistance et la qualité de la représentation des articles. Notons que le corpus de notre collection-test comporte environ 4,5 millions de notices bibliographiques dont 78 % possèdent un résumé et 99 % contiennent entre 10 à 12 descripteurs MeSH associés manuellement par des indexeurs humains. Actuellement la NLM[17], gestionnaire de la base MEDLINE, est confrontée à un réel dilemme. Face au nombre croissant d'articles produits dans le domaine biomédical, l'indexation automatique est attractive et représente un intérêt financier considérable. Dans ce contexte, on s'interroge sur l'efficacité relative de ces deux approches d'indexation en termes de qualité de recherche. Dans cette optique, nous avons évalué ces deux approches. Plus précisément, pour l'indexation automatique nous avons considéré uniquement le titre et le résumé (voir l'exemple dans l'annexe E) tandis que pour l'indexation manuelle nous avons également pris en compte les descripteurs MeSH. Pour consolider cette évaluation, nous avons réalisé cette expérience en utilisant sept modèles de recherche dont les trois probabilistes mentionnés plus haut.

---

[16] MeSH est un vocabulaire contrôlé développé et maintenue par la NLM[17]. Ce thesaurus est organisé de manière hiérarchique et comprend quelques 22'997 descripteurs allant de termes les plus généraux aux termes les plus spécifiques. Il est utilisé pour l'indexation et la recherche dans la base MEDLINE.

[17] *National Library of Medecine* : http://www.nlm.nih.gov/

Enfin, le dernier aspect de cette contribution était consacré à l'expansion de requête. L'objectif était de vérifier si cette technique permettrait d'améliorer la qualité de recherche dans un contexte spécifique d'une part et d'autre part, comparer les performances de l'approche que nous avons introduite dans la section 2.1 par rapport à la méthode de Rocchio (Buckley *et al.*, 1996). Pour réaliser cette expérience, nous avons choisi le modèle I(n)B2.

L'évaluation de diverses stratégies de recherche révèle que le modèle I(n)B2 de DFR permet une amélioration de 174 % par rapport à une approche vectorielle classique $tf \cdot idf$. Si l'on analyse l'impact des descripteurs manuellement ajoutés à chaque notice bibliographique, leur influence s'avère bénéfique quel que soit le modèle de recherche utilisé. Les améliorations apportées par l'indexation manuelle varient entre 2,4 % et 13,5 %. De plus, les tests statistiques indiquent une différence de performance statistiquement significative, et ce en particulier avec les modèles les plus performants. Par ailleurs, les résultats obtenus avec l'expansion de requête étaient quelque peu surprenants. En effet, tandis que l'approche de Rocchio a significativement détérioré la performance, l'approche IDFQE que nous avons proposée a permis d'améliorer la qualité de recherche. Cependant, cette amélioration n'est statistiquement significative qu'avec des paramètres précis, à savoir en sélectionnant 10 termes parmi les 10 premiers documents.

2.4. *Considérations sur l'évaluation de la robustesse en recherche d'information*

Les récents progrès réalisés dans le domaine de la RI, notamment grâce aux diverses campagnes d'évaluation, a permis d'identifier de nouveaux défis. Par exemple, le problème des requêtes difficiles représente un enjeu majeur, et en particulier pour les systèmes commerciaux. En effet, répondre à tous les besoins d'informations de l'usager sans paraître insensé ou stupide face à des requêtes difficiles est une exigence qu'un système commercial se doit d'assurer. Dans ce contexte, la piste robuste a été initiée la première fois en 2003 à TREC (Voorhees, 2004) et en 2006 à CLEF. D'une part, l'objectif principal de cette piste était de focaliser les efforts sur les améliorations apportées sur des requêtes ayant des performances faibles. D'autre part, on cherchait à déterminer si pour les requêtes signalées comme difficiles avec les systèmes utilisés lors des éditions précédentes de TREC, elles demeurent ardues malgré les progrès enregistrés en RI. Finalement, même les modèles de recherche les plus récents peinent à fournir des résultats satisfaisants face à ce type de requêtes. Par ailleurs, d'autres questions comme la définition de la notion de « difficile » pour une requête ou comment évaluer la robustesse d'un système du point de vue efficacité de recherche préoccupaient les participants à cette piste robuste.

L'évaluation des systèmes de RI repose sur un ensemble riche de mesures permettant d'apprécier divers aspects et propriétés de ces systèmes. Parmi ces métriques, la moyenne des précisions moyennes (MAP, *Mean Average Precision*) est largement utilisée pour comparer des systèmes de recherche *ad hoc*. Elle constitue

la mesure officielle dans les conférences TREC, NTCIR et CLEF. Selon l'aspect que l'on désire évaluer, on peut recourir à d'autres métriques comme la précision après cinq (P@5) ou dix (P@10) documents retrouvés, ou la moyenne de l'inverse du rang de la première bonne réponse (MRR, *Mean Reciprocal Rank*). Cette dernière est utilisée par exemple pour l'évaluation des systèmes Web. Pour la piste robuste, lors de sa seconde édition, la moyenne géométrique des précisions moyennes (GMAP, *Geometric Mean Average Precision*) a été proposée afin d'accorder plus d'importance aux améliorations apportées sur des petites performances. Plus récemment, la moyenne du score pondéré du premier document pertinent (FRS, *First Relevant Score*) a été proposée par Tomlinson (2006) pour mesurer la capacité d'un système à retourner une bonne réponse parmi les premières de la liste.

Dans cet article, nous avons abordé les avantages et inconvénients de chacune de ces métriques. En particulier, nous avons comparé trois mesures d'évaluation, à savoir la MAP, la GMAP et la FRS. Pour cela, nous avons évalué six modèles de recherche allant du modèle vectoriel classique $tf \cdot idf$ au plus récent développement dans le paradigme DFR avec le modèle DLH (Amati, 2006) ou le modèle I(n_e)C2 conçu pour apporter de bonnes réponses face à des requêtes difficiles (Plachouras *et al.*, 2005). En comparant le classement de ces approches selon différentes mesures, notre objectif était d'une part, de mettre en évidence les requêtes pour lesquelles les stratégies de recherche rencontreraient des difficultés et d'autre part, déterminer comment évaluer la robustesse d'un système de recherche. Enfin, notre dernière expérience dans cet article évalue la pseudo-rétroaction de requêtes (Buckley *et al.*, 1996) selon ces trois mesures. Nous avons considéré pour cette expérience les modèles Okapi et I(n_e)C2. Toutes les évaluations ont été réalisées sur la collection-test française de la piste robuste de CLEF 2006.

Enfin, nous avons constaté que le classement des modèles peut changer selon qu'on adopte la mesure MAP ou la GMAP. Cette dernière met clairement en avant le modèle « Lnu » (Singhal *et al.*, 1996) devant le modèle DLH ou encore le modèle de langue (Hiemstra, 2001). En analysant les requêtes difficiles selon la précision moyenne, nous avons trouvé une explication à un tel comportement du modèle « Lnu » lorsqu'on considère la GMAP. Ces premiers résultats confirment que même les modèles les plus récents ne sont pas forcément les plus performants face à certaines requêtes. Le classement des modèles obtenu selon la FRS est identique au classement fournit par la GMAP. Mais l'application de ces trois mesures pour l'évaluation de la pseudo-rétroaction de Rocchio révèle des divergences importantes. Ainsi selon la FRS l'expansion de requête détériore la qualité de recherche. Enfin, ces expériences nous ont également permis de mettre en lumière une typologie des raisons expliquant la difficulté intrinsèque de certaines stratégies de recherche face à certaines requêtes.

3. Conclusion

3.1. *Contributions*

Dans cette thèse, nous avons présenté nos travaux en recherche d'information selon deux contextes. En premier lieu, nous nous sommes intéressé au caractère plurilingue de la Toile en abordant le développement de moteurs de recherche pour des langues présentant des caractéristiques visuelles, morphologiques et syntaxiques fort différentes des langues indo-européennes. Plus précisément, nous avons proposé des stratégies de recherche pour les langues chinoise, japonaise, coréenne et, à des fins de comparaison, la langue anglaise. Nous avons utilisé à cet effet des collections-tests comprenant des dépêches d'agences représentant en général le contexte standard de la RI, considéré proche de la réalité du Web puisque la consultation de l'actualité est l'une des raisons les plus importantes de la navigation sur Internet. En second lieu, nous avons focalisé nos efforts dans un contexte spécifique, à savoir le domaine de la biomédecine. Nous avons présenté et adapté des stratégies permettant d'améliorer l'efficacité de nos moteurs de recherche pour ce domaine. De manière générale, nous pouvons tirer les conclusions suivantes.

Premièrement, nous avons constaté que les modèles issus du paradigme *Divergence from Randomness* (Amati & van Rijsbergen, 2002) offrent très souvent des performances meilleures que les modèles vectoriels ou le modèle de langue de Hiemstra (2001). Les différences observées avec le modèle Okapi (Robertson *et al.*, 2000) n'ont pas été toujours statistiquement significatives. Sur la base de nos évaluations, nous pouvons donc affirmer que les modèles de recherche probabilistes paramétriques constituent les stratégies de dépistage les plus efficaces.

Deuxièmement et contrairement à la langue anglaise ou la plupart des langues indo-européennes, le choix de la forme d'indexation est crucial pour les langues chinoise, japonaise et coréenne. En présence de requêtes courtes et pour la langue chinoise, nous avons montré par rapport à une stratégie par bigrammes qu'une combinaison de caractères (ou unigrammes) avec des bigrammes permet une amélioration d'environ 10 % en moyenne tandis qu'une indexation par mots n'apporte que 5 % d'amélioration en moyenne. Une différence statistiquement significative a été plus souvent détectée dans le premier cas (« unigrammes & bigrammes » *vs.* « bigrammes ») que dans le deuxième (« mots » *vs.* « bigrammes »). Pour la langue coréenne, la stratégie par bigrammes permet d'atteindre les meilleurs résultats : l'indexation par morphèmes dégrade la performance d'environ 7 % tandis qu'une approche par mot ignorant tout traitement morphologique occasionne une perte de la performance d'environ 37 % par rapport à une approche par bigrammes. Notons cependant que la différence entre l'approche par bigrammes et la stratégie par morphèmes n'a pas toujours été statistiquement significative. Enfin pour le japonais, l'évidence statistique n'a pas été détectée pour départager l'approche combinée « unigrammes & bigrammes » et la stratégie par

mots. Ces deux méthodes apportent un accroissement de la performance d'environ 7 % par rapport à une indexation par bigrammes.

Troisièmement, nous avons proposé une nouvelle approche pour la reformulation de requêtes afin d'améliorer la performance d'un système de RI. Cette stratégie, nommée « IDFQE », est basée sur la valeur idf de chaque terme ainsi que sa fréquence dans l'ensemble des documents supposés pertinents. Comparée à l'approche de Rocchio (Buckley *et al.*, 1996), nos évaluations montrent que notre algorithme s'est avéré plus souvent meilleur. En particulier, l'application de cette stratégie dans le contexte biomédical, à savoir sur la collection-test MEDLINE de TREC-2004, a montré une amélioration statistiquement significative tandis que l'approche de Rocchio a statistiquement détérioré la performance.

Quatrièmement, en se limitant au domaine spécifique de la biomédecine, nous avons constaté que l'indexation des descripteurs manuellement attribués en plus du titre et du résumé des notices bibliographiques permet des améliorations allant de 2,4 % à 13,5 % suivant le modèle de recherche adopté. Cette progression est particulièrement marquée et statistiquement significative avec les modèles de recherche les plus performants.

Enfin, l'analyse critique de l'évaluation des stratégies de recherche a révélé l'importance de considérer diverses mesures de performance afin de mettre en lumière des phénomènes difficiles à détecter avec une seule métrique. Ainsi, en utilisant la mesure FRS (Tomlinson, 2006), nous avons démontré à l'aide d'une collection-test en langue française que la pseudo-rétroaction de Rocchio (Buckley *et al.*, 1996) cause une détérioration du rang moyen du premier document pertinent. Ce fait explique également pourquoi cette technique de rétroaction n'a pas d'intérêt majeur à être incluse dans les moteurs de recherche commerciaux sur le Web où les premières réponses de la liste des résultats sont très importantes.

3.2. *Perspectives*

Les résultats que nous venons de résumer soulèvent à leur tour de multiples interrogations dont les réponses pourraient améliorer la qualité des moteurs de recherche. Nous présentons dans les paragraphes suivants quelques aspects parmi d'autres auxquels nous nous attacherons à trouver des réponses dans des travaux futurs.

Modèles de recherche

Au cours de nos expériences, nous avons implémenté et évalué de nombreux modèles de recherche. En analysant les performances individuelles des requêtes obtenues avec ces diverses stratégies de recherche, nous avons remarqué que les modèles ne sont pas tous aussi performants face aux différentes requêtes. Nous avons constaté que même les stratégies classiques (par exemple un simple modèle $tf \cdot idf$) peuvent être les plus performantes pour certaines requêtes. Dès lors, il

serait intéressant de développer des méthodes permettant de prédire pour une requête le modèle de recherche le plus efficace. A cet effet, He & Ounis (2004) ont proposé une approche permettant de sélectionner de façon optimale le modèle à appliquer pour une requête donnée. Dans cette approche, les requêtes de l'ensemble d'apprentissage sont d'abord regroupées dans des clusters en fonction de leurs caractéristiques statistiques et du modèle de recherche le plus performant pour le cluster. Ensuite pour une nouvelle requête, la sélection de modèle se fait en calculant la distance du cluster le plus proche à cette requête. Enfin, les résultats obtenus avec cette approche de sélection de modèle n'ont pas dépassé la performance obtenue avec le meilleur modèle unique. Une des approches que nous pouvons suggérer pour résoudre ce problème est l'utilisation de la régression logistique pour prédire le modèle à utiliser.

Par ailleurs, nous avons constaté que le modèle Okapi (Robertson *et al.*, 2000) offre des performance relativement stables indépendamment de la collection-test, de la langue, de la longueur des requêtes ou de la forme d'indexation. Par conséquent, face aux bonnes performances obtenues avec certaines approches du paradigme DFR (Amati & van Rijsbergen, 2002 ; Amati, 2006), il devient nécessaire de déterminer s'il existe parmi l'ensemble des modèles que l'on peut dériver de ce paradigme une approche particulière qui permet de fournir des résultats performants et stables dans diverses situations. Ce problème se pose également avec les modèles de langue. De plus, les résultats que nous avons obtenus avec le modèle de Hiemstra (2001) sont relativement décevants, en particulier pour les trois langues asiatiques. Dans ce cas, l'évaluation d'autres approches de lissage ainsi que d'autres modèles de langue (Zhai & Lafferty, 2004) nous permettrait d'une part de comprendre les raisons de cet échec et d'autre part de déterminer s'il existe un modèle de langue performant indépendamment de la collection-test, de la langue, de la longueur des requêtes ou d'autres critères.

Pseudo-rétroaction

La plupart des approches d'expansion de requêtes, y compris notre stratégie IDFQE, reposent sur un ensemble de paramètres, dont certains sont importants comme le nombre de documents et de termes à considérer pour construire les nouvelles requêtes. Ces paramètres sont souvent déterminés de manière empirique et restent identiques pour l'ensemble des requêtes. Dans ce cas, nous sommes amenés à nous demander s'il existe un moyen d'optimiser ce choix individuellement pour chaque requête. Pour cela, il conviendrait éventuellement d'exploiter les informations statistiques des requêtes et des documents dépistés.

Par ailleurs, l'expansion de requêtes selon notre approche IDFQE peut dans certains cas complètement modifier la requête initiale. Dès lors, nous nous interrogeons sur la différence qu'il peut y avoir entre l'ajout de termes supplémentaires à la requête de départ ou sa reformulation complète. Dans ces cas, comment pondérer les termes est une des questions à laquelle nous devons trouver une réponse plus efficace.

4. Références

Abdou, S., Savoy, J. (2005). Report on CLIR Task for the NTCIR-5 Evaluation Campaign. In *Proceedings of the 5th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access* (pp. 13-24). Tokyo, Japan: National Institute of Informatics.

Abdou, S., Ruch, P., & Savoy, J. (2006). Evaluation of Stemming, Query Expansion and Manual Indexing Approaches. In E. M. Voorhees, & L. P. Buckland (Eds.), *Proceedings of the Fourteenth Text Retrieval Conference (TREC-2005)* NIST Special Publication 500-266. Available at http://trec.nist.gov/pubs/trec14/t14_proceedings.html (last visited, May 1st, 2006).

Abdou, S., Savoy, J. (2007). Monolingual Experiments with Far-East Languages in NTCIR-6. In *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access* (pp. 52-59). Tokyo, Japan: National Institute of Informatics.

Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., Dumais, S., Fuhr, N., Harman, D. K., Harper, D. J., Hiemstra, D., Hofmann, T., Hovy, E., Kraaij, W., Lafferty, J., Lavrenko, V., Lewis, D., Liddy, L., Manmatha, R., McCallum, A., Ponte, J., Prager, J., Radev, D., Resnik, P., Robertson, S., Rosenfeld, R., Roukos, S., Sanderson, M., Schwartz, R., Singhal, A., Smeaton, A., Turtle, H., Voorhees, E. M., Weischedel, R., Xu, J., & Zhai, C. (2003). Challenges in Information Retrieval and Language Modeling: Report of a Workshop Held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002. *SIGIR Forum, 37* (1), 31-47.

Amati, G., & van Rijsbergen, C. J. (2002). Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. *ACM Transaction on Information Systems (TOIS)*, *20* (4), 357-389.

Amati, G. (2006). Frequentist and Bayesian Approach to Information Retrieval. In M. Lalmas, A. MacFarlane, S. M. Rüger, A. Tombros, T. Tsikrika, & A. Yavlinsky (Eds.), *Advances in Information Retrieval, 28th European Conference on IR Research (ECIR'06)* (pp. 13-24), LNCS, vol. 3936. Berlin, Germany: Springer.

Bergman, M. K. (2001). The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing, 7* (1). Available at http://www.press.umich.edu/jep/07-01/bergman.html (last visited, May 1st 2007).

Braschler, M., & Peters, C. (2004). Cross-Language Evaluation Forum: Objectives, Results, Achievements. *Information Retrieval, 7* (1/2), 7-31.

Braschler, M., & Ripplinger, B. (2004). How Effective is Stemming and Decompounding for German Text Retrieval?. *Information Retrieval, 7* (3/4), 291-316.

Buckley, C., Singhal, A., Mitra, M., & Salton, G. (1996). New Retrieval Approaches using SMART. In D. K. Harman (Ed.), *Proceedings of the Fourth Text Retrieval Conference (TREC-4)*, NIST Special Publication 500-236 (pp. 25-48). Gaithersburg, MS, USA: NIST.

Chen, A. (2003). Cross-Language Retrieval Experiments at CLEF-2002. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Advances in Cross-Language Information Retrieval, Third Workshop of Cross-Language Evaluation Forum, CLEF 2002, Rome, Italy, September 19-20 2002, Revised Papers* (pp. 28-48). LNCS, vol. 2785. Berlin, Germany: Springer.

Chen, A., & Gey, F. C. (2003). Experiments on Cross-Language and Patent Retrieval at NTCIR-3 Workshop. In *Proceedings the Third NTCIR Workshop Meeting on Evaluation of Information Access Technologies* (p. 216-224). Tokyo, Japan: National Institute of Informatics.

Chung, T. L., Luk, R. W. P., Wang, K. F., Kwok, K. L., & Lee, D. L. (2006). Adapting Pivoted Document-Length Normalization for Query Size: Experiments in Chinese and English. *ACM Transactions on Information Systems (TOIS), 5* (3), 245-263.

Gordon, M., & Kochen M. (1989). Recall-precision trade-off: A derivation. *Journal of the American Society for Information Science (JASIS), 40* (3), 145-151.

Gospodnetić, O., & Hatcher, E. (2005). *Lucene in Action*. Greenwich, CT, USA: Manning Publications Co.

Grefenstette, G. (1998). *Cross-Language Information Retrieval*. Norwell, MA, USA: Kluwer Academic Publishers.

Harman, D. K. (1995). Overview of the Third Text REtrieval Conference (TREC-3). In D. K. Harman (Ed.), *Proceedings of the Third Text Retrieval Conference (TREC-3)* (NIST Special Publication 500-225, pp. 1-20). Gaithersburg, MS, USA: NIST.

Harman, D. K. (2005a). Beyond English. In E. M. Voorhees, & D. K. Harman (Eds.), *TREC Experiment and Evaluation in Information Retrieval* (pp. 153-181). Cambridge, Massachusetts, USA: The MIT Press.

Harman, D. K. (2005b). The History of IDF and its Influences on IR and Other Fields. In J. I. Tait (Ed.), *Charting a New Course: Natural Language Processing and Information Retrieval Essays in Honor of Karen Sparck Jones* (pp. 69-79). Dordrecht, The Netherlands: Springer.

He, B., & Ounis, I. (2004). A Query-based Pre-retrieval Model Selection Approach to Information Retrieval. In *Proceedings of RIAO 2004 (Recherche d'Information Assistee par Ordinateur)* (pp. 709-719).

He, B., & Ounis, I. (2005). A Study of Parameter Tuning for Term Frequency Normalization. In *Proceedings of the twelfth international conference on Information and knowledge management (CIKM'03)* (pp. 10-16), New York, USA: ACM Press.

Hersh, W. R., Bhuptiraju, R. T., Ross, L., Johnson, P., Cohen, A. H. & Kraemer, D. F. (2005). TREC 2004 Genomics Track Overview. In E. M. Voorhees, & L. P. Buckland (Eds.), *Proceedings of the Thirteenth Text Retrieval Conference (TREC-2004)*, NIST Special Publication 500-261 (pp. 192-201). Gaithersburg, MS, USA: NIST.

Hiemstra, D. (1998). A Linguistically motivated probabilistic Model of Information Retrieval. In N. Christos, & C. Stephanides (Eds.), *Proceedings of European Conference of Digital Library (ECDL'98)* (pp. 569-584). LNCS, vol. 1513. London, UK: Springer.

Hiemstra, D. (2001). Using Language Models for Information Retrieval. PhD thesis, University of Twente.

Hollink, V., Kamps, J., Monz, C., & de Rijke, M. (2004). Monolingual Document Retrieval for European Languages. *Information Retrieval, 7* (1/2), 33-52.

Ihadjadene, M. (2004). *Les systèmes de recherche d'informations*. Paris, France: Lavoisier.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real Life, Real Users, and Real Needs: a Study and Analysis of User Queries on the Web. *Information Processing and Management, 36* (2), 207-227.

Jansen, B. J., & Spink, A. (2006). How are we Searching the World Wide Web? A Comparison of Nine Search Engine Transaction Logs. *Information Processing and Management, 42* (1), 248-263.

Jones, G. J. F. (2005). Beyond English Text: Multilingual and Multimedia Information Retrieval. In J. I. Tait (Ed.), *Charting a New Course: Natural Language Processing and Information Retrieval Essays in Honor of Karen Sparck Jones* (pp. 81-97). Dordrecht, The Netherlands: Springer.

Kando, N., & Adachi, J. (2004). Report from NTCIR Workshop 3. *SIGIR Forum, 38* (1), 10-16.

Kishida, K., Chen, K.-H., Lee, S., Chen, H.-H., Kando, N., Kuriyama, K., Myaeng, S. H., & Eguchi K. (2004). Cross-Lingual Information Retrieval (CLIR) Task at the NTCIR Workshop 3. *SIGIR Forum, 38* (1), 17-20.

Kishida, K., Chen, K.-H., Lee, S., Kuriyama, K., Kando, N., Chen, H.-H., & Myeng, S.-H. (2005). Overview of CLIR Task at the Fifth NTCIR Workshop. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access* (pp. 1-10). Tokyo, Japan: National Institute of Informatics

Kishida, K., Chen, K.-H., Lee, S., Kuriyama, K., Kando, N., & Chen, H.-H. (2007). Overview of CLIR Task at the Sixth NTCIR Workshop. In *Proceedings of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access* (pp. 1-19). Tokyo, Japan: National Institute of Informatics

Kwok, K. L. (1999). Employing Multiple Representations for Chinese Information Retrieval. *Journal of the American Society for Information Science (JASIS), 50* (8), 709-723.

Kwok, K. L., Dinstl, N., & Choi, S. (2004). NTCIR-4 Chinese, English, Korean Cross Language Retrieval Experiments using PIRCS. In *Proceedings the Fourth NTCIR Workshop Meeting on Evaluation of Information Access Technologies* (pp. 186-192). Tokyo, Japan: National Institute of Informatics.

Lee J. H., & Ahn J. S. (1996). Using n-grams for Korean Text Retrieval. In H.-P. Frei, D. Harman, P. Schäuble, & R. Wilkinson (Eds.), *Proceedings of the 19th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 216-224), New-York, USA: ACM Press.

Luk, R. W. P., Kwok, K. L. (2002). A Comparison of Chinese Document Indexing Strategies and Retrieval Models. *ACM Transaction on Asian Language and Information Processing (TALIP), 1* (3), 224-268.

Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., & Asahara, M. (1999). Japanese Morphological Analysis System ChaSen. Technical Report NAIST-IS-TR99009. Available at http://chasen.aist-nara.ac.jp (last visited, May 1st 2007).

McNamee, P. (2001). Experiments in the Retrieval of Unsegmented Japanese Text at the NTCIR-2 Workshop. In *Proceedings the Second NTCIR Workshop Meeting on Evaluation of Information Access Technologies* (pp. 157-162). Tokyo, Japan: National Institute of Informatics.

McNamee, P., & Mayfield, J. (2003). Scalable Multilingual Information Access. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Advances in Cross-Language Information Retrieval, Third Workshop of Cross-Language Evaluation Forum, CLEF 2002, Revised Papers* (pp. 207-218). LNCS, vol. 2785. Berlin, Germany: Springer.

McNamee, P., & Mayfield, J. (2004). Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval, 7* (1/2), 73-97.

Miller, D. R. H., Leek, T., & Schwartz, R. M. (1999). A Hidden Markov Model Information Retrieval System. In M. Agosti, & M. Melucci (Eds.), *Proceedings of the 22nd Annual international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 214-221), New York, USA: ACM Press.

Nakagawa, H., Mori, T., & Kando, N. (Eds.). (2005). Preface to Special Issues on NTCIR-4. *ACM Transactions on Asian Language Information Processing (TALIP), 4* (3), 237-242.

Nie, J.-Y., & Ren F. (1999). Chinese Information Retrieval: Using Characters or Words?. *Information Processing & Management (IP&M), 35* (4), 443-462.

Nie, J.-Y., Gao, J., Zhang, J., & Zhou, M. (2000). On the Use of Words and N-grams for Chinese Information Retrieval. In *Proceedings IRAL* (pp. 141-148), New York, USA: ACM Press.

Oard, D. W., & Resnik, P. (1999). Support for Interactive Document Selection in Cross-Language Information Retrieval. *Information Processing and Management, 35* (4), 363-379.

Peters, C., Clough, P., Gonzalo, J., Jones, G. J. F., Kluck, M., Magnini, B., & de Rijke, M. (Eds.). (2005). *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Revised Selected Papers*. LNCS, vol. 3491. Berlin, Germany: Springer.

Peters, C., Gey, F. C., Gonzalo, J., Müller, H., Jones, G. J. F., Kluck, M., Magnini, B., & de Rijke, M. (Eds.). (2006). *Accessing Multilingual Information Repositories, 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Revised Selected Papers*. LNCS, vol. 4022. Berlin, Germany: Springer.

Pirkola, A. (2001). Morphological Typology of Languages for IR. *Journal of Documentation, 57* (3), 330-348.

Plachouras, V., He, B., & Ounis, I. (2005). University of Glasgow at TREC 2004: Experiments in web, robust and terabytes tracks with Terrier. In E. M. Voorhees, & L. P. Buckland (Eds.), *Proceedings of the Fourteenth Text Retrieval Conference (TREC-2005)* NIST Special Publication 500-261. Available at http://trec.nist.gov/pubs/trec14/t14_proceedings.html (last visited, May 1st, 2006).

Ponte, J., & Croft, B. (1998). A Language Modeling Approach in Information Retrieval. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 275-281), New York, USA: ACM Press.

Robertson, S. E. (1977). The Probability Principle Ranking in IR. *Journal of Documentation, 33* (4), 294-304.

Robertson, S. E., & Walker, S. (1994). Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In W. B. Croft, & C. J. van Rijsbergen (Eds.), *Proceedings of the 17th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 232-241), ACM Press/Springer.

Robertson, S. E., Walker, S., & Beaulieu, M. (2000). Experimentation as a Way of Life: Okapi at TREC. *Information Processing and Management*, *36* (1), 2000, 95-108.

Robertson, S. E. (2004). Understanding Inverse Document Frequency: On theoretical arguments of IDF. *Journal of Documentation*, *60* (5), 503-520.

Salton, G. (1968). *Automatic Information Organisation and Retrieval.* New-York, USA: McGraw-Hill.

Salton, G. (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*. New Jersey, USA: Prentice-Hall.

Salton ,G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New-York, USA: McGraw-Hill.

Saracevic, T. (1975). Relevance: A Review of the Literature and a Framework for the Thinking on the Notion in Information Science. *Journal of the American Society for Information Science, 25* (6)*,* 321-343.

Saracevic, T. (2006). Relevance: A Review of the Literature and a Framework for the Thinking on the Notion in Information Science. Part II. *Advances in Librarianship, 30*, 3-71.

Savoy, J. (2002). Recherche d'information dans des corpus plurilingues. *Ingénierie des systèmes d'informations, 7* (1-2), 63-93.

Savoy, J. (2003). Report on CLEF-2002 Experiments: Combining Multiple Sources of Evidence. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Advances in Cross-Language Information Retrieval, Third Workshop of Cross-Language Evaluation Forum, CLEF 2002, Revised Papers* (pp. 66-90). LNCS, vol. 2785. Berlin, Germany: Springer.

Savoy, J. (2005). Comparative Study of Monolingual and Multilingual Search Models for Use with Asian Languages. *ACM Transaction on Asian Language and Information Processing (TALIP), 4* (2), 163-189.

Savoy, J., & Abdou, S. (2006). UniNE at CLEF 2006: Experiments with Monolingual, Bilingual, Domain-Specific and Robust Retrieval. In *Working Notes of the 6th Workshop on Cross-Language Evaluation Forum, CLEF 2006.* Available at http://www.clef-campaign.org/2006/working_notes/CLEF2006WN-Contents.html (last visited, May 1st, 2006).

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted Document Length Normalization. In H.-P. Frei, D. Harman, P. Schäuble, & R. Wilkinson (Eds.), *Proceedings of the 19th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 21-29), New-York, USA: ACM Press.

Song, F., & Croft, W. B. (1999). A General Language Model for Information Retrieval. In *Proceedings of the Eighth International Conference on Information and Knowledge Management (CIKM'99)* (pp. 316-321). New-York, USA: ACM Press.

Sparck-Jones, K. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, *28* (1), 11-21. (reprinted in *Journal of Documentation*, *50* (5), 2004, 493-502).

Sparck-Jones, K., Walker, S., & Robertson, S. E. (2000a). A Probabilistic Model of Information Retrieval: Development and Comparative Experiments - Part 1. *Information Processing and Management*, *36* (6), 779-808.

Sparck-Jones, K., Walker, S., & Robertson, S. E. (2000b). A Probabilistic Model of Information Retrieval: Development and Comparative Experiments - Part 2. *Information Processing and Management*, 36 (6), 809-840.

Tomlinson, S. (2004). Experiments with Decompounded Chinese, Japanese and Korean Words by Hummingbird™ SearchSearver at NTCIR-4. In *Proceedings the Fourth NTCIR Workshop Meeting on Evaluation of Information Access Technologies* (pp. 128-135). Tokyo, Japan: National Institute of Informatics.

Tomlinson, S. (2004). Bulgarian and Hungarian Experiments with Hummingbird™ SearchServer at CLEF 2005. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini & M. de Rijke (Eds.), *Accessing Multilingual Information Repositories, Sixth Workshop of Cross-Language Evaluation Forum, CLEF 2005, Revised Papers* (pp. 194-203). LNCS, vol. 4022. Berlin, Germany: Springer.

Van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). London, UK: Butterworth.

Voorhees, E. M. (1998). Variations in Relevance Judgments and the Measurements of Retrieval Effectiveness. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 315-323), New York, USA: ACM Press.

Voorhees, E. M. (2002). The Philosophy of Information Retrieval Evaluation. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001, Revised Papers* (pp. 355-379). LNCS, vol. 2406. Berlin, Germany: Springer.

Voorhees, E. M.. (2004). Overview of the TREC 2003 Robust Retrieval Track. In E. M. Voorhees, & L. P. Buckland (Eds.), *Proceedings of the Twelfth Text Retrieval Conference*

(TREC-2003) NIST Special Publication 500-255. Available at http://trec.nist.gov/pubs/trec12/t12_proceedings.html (last visited, May 1st, 2006).

Voorhees, E. M., & Harman, D. K. (Eds.). (2005). *TREC Experiment and Evaluation in Information Retrieval*. Cambridge, Massachusetts, USA: The MIT Press.

Weeber, M., Schijvenaars, B. J. A., Van Mulligen, E. M., Mons, B., Jelier, R., Van Der Eijk, C., & Kors, J. A. (2003). Ambiguity of Human Gene Symbols in LocusLink and MEDLINE: Creating an Inventory and a Disambiguation Test Collection, In M. A. Musen (Ed.), *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium* (pp. 704-708). Philadelphia, PA, USA: Hanley & Belfus, Inc.

Wei, J. (2004). Worldwide Internet Usages and Online Multi-linguistic Population Comparison Study. *Information Systems Education Journal, 2* (25), 239-251.

Yu, H., Kim, W., Hatzivassiloglou, V., & Wilbur, J. (2006). A Large Scale, Corpus-Based Approach for Automatically Disambiguating Biomedical Abbreviations. *ACM Transactions on Information Systems (TOIS), 24* (3), 380-404.

Zhai, C., & Lafferty, J. (2004). A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems (TOIS), 22* (2), 179-214.

Annexes

A. ***Exemple d'un document en langue chinoise (NTCIR-5)***

<DOC>
<DOCNO>ude_xxx_20000101_0264850</DOCNO>
<LANG>CH</LANG>
<HEADLINE>龍象 除夕練兵職籃 活力再現 - 記者王樹衡／台北報導 -
千禧籃賽今開打</HEADLINE>
<DATE>2000-01-01</DATE>
<TEXT>
<P>「新世紀、新希望」為職籃復賽熱身，也是千禧年國內首項體育盛事的職籃千禧紀念賽今天點燃戰火，首戰由宏國象、裕隆恐龍兩隊下午4時30分在台北體院體育館對決，一直關心職籃復賽的副總統連戰將蒞臨現場，為球員打氣。職籃自去年3月封館以來，造成球員流失，鄭志龍遠赴上海打工就是一例，而經過各界努力，終於讓職籃在千禧年第一天重燃戰火，並將在4月正式復賽，展開職籃六年球季，千禧年對職籃而言將是重新出發的一年。自職籃元年開打以來，宏國、裕隆均扮演開幕的角色，千禧賽也不例外，龍象大戰戲碼將喚起球迷的回憶，兩隊都在昨天除夕繼續練球，準備全力在千禧年爭取好采頭。獲職籃二年、三年、四年三季總冠軍的宏國隊，主力陣容少了鄭志龍、朱浩仁，但加入不少年輕新秀如歐陽敬恆、李豐永等，速度增加不少，另外周俊三、邱德治、羅興樑、黃春雄、劉義祥，以及洋將「台灣女婿」雷克斯仍在陣中，戰力仍屬一級。裕隆隊這一年休兵期產生極大變化，改以年輕化為取向，網羅業餘最佳球員陳信安配合李雲光、林建平、東方介德、邱宗志、蔡福財等老將，加上洋將豪爾前天歸隊，與宏國仍有得拚。下午2時30分起擔任主場的裕隆隊將在台北體院體育館前舉行年輕球員陳信安、邱啟益等人簽名會先行造勢。</P>
</TEXT>
</DOC>

B. ***Exemple d'un document en langue japonaise (NTCIR-5)***

<DOC>
<DOCNO>JA-000101047</DOCNO>
<LANG>JA</LANG>
<SECTION>経済</SECTION>
<AE>無</AE>
<WORDS>534</WORDS>
<HEADLINE>原油価格、３５％高－－１９９９年</HEADLINE>
<DATE>2000-01-01</DATE>
<TEXT>
１９９９年の原油の平均価格は、前年に比べ３５％上昇したことが３１日までに、明らかになった。石油輸出国機構（ＯＰＥＣ）諸国などによる原油の協調減産が影響した。ＯＰＥＣは現在の減産期限である今年３月末以降も減産を維持する構えで、２０００年も原油価格が高水準のまま推移すれば、世界的なインフレの原因になる可能性もある。
１２月３０日で昨年の取引を終えたロンドン国際石油取引所によると、指標銘柄の北海ブレント先物の昨年の平均価格は１バレル＝１８ドルだった。９８年の平均価格は１バレル＝１３・３４ドルで、４ドル以上も上昇した。
値上がりの最大の要因は、ＯＰＥＣとメキシコなど非ＯＰＥＣ産油国による日量５００万バレルを超える協調減産。米国経済が好調を維持し、アジア経済も順調に回復して原油の需要が回復したことが、価格を押し上げた。
ただ、原油需給のひっ迫は、ＯＰＥＣなどの予想を上回る形で進んでおり、国際エネルギー機関（ＩＥＡ）は「今冬、一時的な原油不足が生じる可能性もある」と指摘している。このため、ＯＰＥＣが３月の総会で協調減産の規模を縮小しなければ「原油価格は高止まりし、工業製品の値上がりを誘発する可能性も高い」（市場関係者）との見方も出始めている。【ロンドン・松木健】
</TEXT>
</DOC>

C. ***Exemple d'un document en langue coréenne (NTCIR-5)***

```
<DOC>
<DOCNO>CHOSUN2000_00006</DOCNO>
<LANG>KR</LANG>
<HEADLINE>공무원 채용때 장애인 5%로 내년7월 시행;신장질환자등
14만명 추가;</HEADLINE>
<DATE>20000101</DATE>
<TEXT>
공무원 채용때 장애인 5%로 내년7월 시행신장질환자등 14만명 추가
내년부터 각종 조세감면과 이용료 면제 혜택을 받을 수 있는 장애인의
범주에 만성 심장, 신장 질환자 등도 포함된다.
또 국가 또는 지방자치단체가 의무적으로 신규 채용해야 하는 장애인
비율이 현행 2%에서 5%로 확대된다.
26일 기획예산처와 보건복지부에 따르면 현재 신체 및 정신지체, 언어,
청각, 시각 등으로 정해져 있는 장애인 범주를 만성심장, 신장질환자,
정신질환자(1급)까지로 확대해 장애인으로서 각종 수혜를 받을 수
있도록 할 계획이라고 밝혔다.
이렇게 될 경우 등록 장애인 수는 현재 64만명에서 78만명선으로 늘어날
전망이다.
또 국가 및 지자체의 경우 장애인 공무원(현재 3600명)이 1만명에 달할
때까지 한시적으로 공무원 신규채용시 장애인 의무채용 비율을 2%에서
5%로 높여 내년 7월부터 시행한다.
 이와 함께 장애인 고용 비율이 1% 미만인 업체가 물어야 하는
고용부담금(미달된 장애인수×1인당 최저임금의 60%)을 최저임금의
70%(25만3000원)로 인상키로
했다.
/이준기자 junlee@chosun.com
</TEXT>
</DOC>
```

D. ***Exemple d'un document en langue anglaise (NTCIR-5)***

```
<DOC>
<DOCNO>ENY-20000101E1TDY03D000020</DOCNO>
<LANG>EN</LANG>
<SECTION>Natl2</SECTION>
<HEADLINE>Y2K danger extends beyond 1st day of
year</HEADLINE>
<DATE>2000-01-01</DATE>
<TEXT>
 Those concerned about Y2K computer malfunctions will
not necessarily be able to heave a sigh of relief once
Jan. 1 has safely passed.
 Several other dates are lurking with a threat to
unleash mayhem in computer systems.
 The Y2K bug could start biting in earnest Tuesday, when
most of the nation's companies return their computer
system to full operations as they resume business.
 Another danger zone is the change of fiscal year when
corporations and government offices settle their
accounts--with March 31 marking the end of the current
year and April 1 the beginning of the new business year.
 To make matters worse, this year marks a special kind
of leap year that falls only once every 400 years.
 A leap year usually falls in years divisible by four,
but years that end in "00"; are not usually treated as
such. However, years that end in "00" and are also
divisible by 400 are treated as regular leap years. Thus
the year 2000 will see 29 days in February.
 It is feared that computer systems that have not taken
this principle into account could break down on Feb. 29.
 The following is a list of red-flag days for potential
Y2K problems:
 -- Jan. 4--the first business day of the year.
 -- Jan. 31--the end of the first month.
 -- Feb. 29--the final day of February in a leap year.
 -- March 31--the end of fiscal 1999.
 -- April 1--the beginning of fiscal 2000.
 -- April 3--the first business day of fiscal 2000.
</TEXT>
</DOC>
```

E. ***Exemple abrégé d'une notice bibliographique extraite de MEDLINE***

```
<DOC>
<PMID>10605453</PMID>
<TI>Immunocytochemical localization studies of myelin
basic protein.</TI>
<AB>The location of myelin encephalitogenic or basic
protein (BP) in peripheral nervous system (PNS) and
central nervous system (CNS) was investigated by
immunofluorescence and horseradish peroxidase (HRP)
immunocytochemistry. BP or cross-reacting material could
be clearly localized to myelin by immunofluorescence and
light microscope HRP immunocytochemistry. Fine tructural
studies proved to be much more difficult, especially in
the CNS, due to problems in tissue fixation and
penetration of reagents. Sequential fixation in aldehyde
followed by ethanol or methanol provided the best
conditions for ultrastructural indirect
immunocytochemical studies. In PNS tissue, anti-BP was
localized exclusively to the intraperiod line of myelin.
Because of limitations in technique, the localization of
BP in CNS myelin could not be unequivocally determined.
In both PNS and CNS tissue, no anti-BP binding to
nonmyelin cellular or membranous elements was
detected.</AB>
<AU>Mendell JR</AU>
<AU>Whitaker JN</AU>
<RN>0 (Myelin Basic Proteins)</RN>
<RN>EC 1.11.1.- (Horseradish Peroxidase)</RN>
<MH>Animals</MH>
<MH>Brain/*cytology/ultrastructure</MH>
<MH>Cattle</MH>
<MH>Caudate Nucleus/cytology/ultrastructure</MH>
<MH>Femoral Nerve/*cytology/ultrastructure</MH>
<MH>Guinea Pigs</MH>
<MH>Haplorhini</MH>
<MH>Horseradish Peroxidase</MH>
<MH>Human</MH>
<MH>Immunohistochemistry/methods</MH>
<MH>Microscopy, Electron</MH>
<MH>Myelin Basic Proteins/*analysis</MH>
<MH>Myelin Sheath/ultrastructure</MH>
<MH>Sciatic Nerve/*cytology/ultrastructure</MH>
<MH>Stellate Ganglion/*cytology/ultrastructure</MH>
…
</DOC>
```

Report on CLIR Task for the NTCIR-5 Evaluation Campaign

Samir ABDOU, Jacques SAVOY
Institut interfacultaire d'informatique, University of Neuchatel
Pierre-à-Mazel 7, 2000 Neuchatel, Switzerland
{ Samir.Abdou, Jacques.Savoy }@unine.ch

Abstract

This paper describes our second participation in an evaluation campaign involving the Chinese, Japanese, Korean and English languages (NTCIR-5). Our participation is motivated by four objectives: 1) study the retrieval performances of various IR models for these languages; 2) compare the relative retrieval effectiveness of bigram and automatic word-segmenting approaches for Chinese and Japanese languages; 3) propose a new blind-query expansion hopefully capable of improving mean average precision; and 4) evaluate the relative performance of the various merging strategies used to combine separate result lists extracted from a corpus written in English, Chinese, Japanese or Korean.

Keywords: CLIR, MLIR, blind-query expansion, probabilistic IR model.

1 Monolingual IR for Asian languages

1.1 Overview of NTCIR-5 test collection

Table 1 displays various statistics from the fifth NTCIR corpora (for more information, see [5]). In this paper, when analyzing the number of pertinent documents per topic, we only considered rigid assessments and thus only "highly relevant" and "relevant" items are seen as being relevant. A comparison of the number of relevant documents per topic, as shown in Table 1, indicated that for the English collection the median number of relevant items per topic was 33, while for the Chinese corpus it was only 26 and 25.5 for the Korean and 24 for the Japanese. Clearly, the number of relevant articles was greater for the English (3,073) corpus, when compared to the Japanese (2,112), Chinese (1,885) or Korean (1,829) collections.

For the various search models used, the bottom part of Table 1 provides an overview their efficiency, indicating the size of each collection in terms of storage space requirements. For example, the row labeled "# postings" indicates the number of indexing terms (words or bigrams) in the inverted file, followed by the size of this inverted file and the time (user CPU time + system CPU time) needed to build it. For the Chinese and Japanese languages we used both the bigram and an automatic word segmentation approach. To carry out our experiments we used a 2 x Intel Xeon 3.06 GHz (memory: 3.6 GB, swap: 15 GB, disk: 5 x 250 GB). The average query size (expressed in number of tokens following stopword removal) and time (in seconds) required to execute both short (T only) and medium-size (DN) queries is shown in the lower rows (computations made without blind-query expansion). Clearly, the use of bigrams as indexing

	English	Chinese		Japanese		Korean
Size (in MB)	438 MB	1,100 MB		1,100 MB		312 MB
# of documents	259,050	901,446		858,400		220,374
# of topics	49	50		47		50
# rel. items	3,073	1,885		2,112		1,829
Mean	62.73	37.7		44.94		36.58
Median	33	26		24		25.5
Indexing scheme	word	word	bigram	word	bigram	bigram
# postings	494,745	333,017	3,661,338	329,884	909,631	345,751
Inverted file size	278 MB	1,786 MB	3,386 MB	955 MB	1,387 MB	586 MB
Building time	1,150 sec.	2,397 sec.	4,726 sec.	1,650 sec.	2,044 sec.	757 sec.
T query size	4.8 wd/que	5.3 wd/que	6.8 bi/que	4.6 wd/que	8.2 bi/que	7.3 bi/que
Search time	0.218 sec	0.275 sec	0.246 sec	0.232 sec	0.270 sec	0.233 sec
DN query size	69.8 wd/que	94.0 wd/que	173.3 bi/que	68.8 wd/que	100.7 bi/que	140.8 bi/que
Search time	0.409 sec	1.631 sec	1.066 sec	0.712 sec	0.770 sec	0.537 sec

Table 1. NTCIR-5 CLIR test collection statistics (rigid evaluation)

strategy required more time to build the inverted file (e.g., for the Chinese corpus, with the time increasing from 2,397 sec. to 4,726 sec., or by 97.2%). The time differences between word-based and bigram searches were not really important.

1.2 Indexing and searching strategies

In analyzing these new test collections and in order to draw some useful conclusions, we considered it important to evaluate the retrieval performance under various conditions. We decided to evaluate a variety of indexing and search models in order to obtain this broader view. First we considered adopting a binary indexing scheme in which each document (or topic) was represented by a set of indexing terms (word or bigram), without assigning any weights (IR model denoted "doc=bnn, query=bnn" or "bnn-bnn"). In order to weight the presence of each indexing term, we might account for the term occurrence frequency ("nnn-nnn") or we might also account for their frequency within the collection (or for *idf*). Moreover, when using cosine normalization, each indexing weight could vary within the range of 0 to 1 ("ntc-ntc" or "*tf idf*").

Other variants might also be created. For example, the *tf* component could be computed as 0.5+0.5·[*tf* / max *tf* in a document] ("atn"). We could also consider that a term's presence in a shorter document provides stronger evidence than in a longer document, leading to more complex IR models; i.e. the IR models denoted by "Lnu" [2] and "dtu" [12]. See the Appendix for details on the exact weighting formulas.

In addition to previous models based on the vector-space model, we also considered probabilistic approaches, such as the well-known Okapi model (or BM25) [8]. As with other probabilistic models, we might apply the Deviation from Randomness (DFR) framework [1], based on two information measures. These are Inf^1 (measuring the informative content of the document with respect to the whole collection), and Inf^2 (measuring the information gain with respect to the *elite* set, the set of documents where the underlying term occurs). To reflect the indexing weight w_{ij} attached to term t_j in document D_i, we have:

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2[Prob^1_{ij}] \cdot (1 - Prob^2_{ij}) \quad (1)$$

in which $Prob^1_{ij}$ is the probability of having by pure chance tf_{ij} occurrences of the term t_j in a document (various probabilistic models could be used to estimate this probability). On the other hand, $Prob^2_{ij}$ is the probability of encountering a new occurrence of term t_j in the given document, once tf_{ij} occurrences of this term have already been found.

Within this DFR framework, the PB2 model is defined as follows:

$$Inf^1_{ij} = -\log_2[(e^{-\lambda_j} \cdot \lambda_j^{tf_{ij}})/tf_{ij}!] \quad (2)$$

$$Prob^2_{ij} = 1 - [(tc_j + 1) / (n \cdot (tfn_{ij} + 1))] \quad (3)$$

with $tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((c \cdot mean\ dl) / l_i)]$

and $\lambda_j = tc_j / n$

where tc_j indicates the number of occurrences of term t_j in the collection, l_i the length (number of indexing terms) of document D_i, *mean dl* the average document length, and *c* a constant.

As a variant, the model denoted I(*n*)L2 (used only for the English corpus) is defined as follows:

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (df_j+0.5)] \quad (4)$$

$$Prob^2_{ij} = tfn_{ij} / (tfn_{ij} + 1) \quad (5)$$

where df_j indicates the number of documents indexed using the term t_j, and *n* the number of documents in the corpus.

In defining these various IR models, we have implicitly admitted that words are our indexing unit. For the English language, finding words in a sentence is usually a simple task. For the Japanese language, each sentence was automatically segmented using the morphological analyzer ChaSen [7], and the Chinese corpus was segmented using Mandarin Tools (freely available at www.mandarintools.com).

For the Asian languages, we also indexed documents by applying an overlapping bigram approach, an indexing scheme found to be effective for various Chinese collections [6], or during previous NTCIR campaigns [3], [11]. Based on this technique for example, the sequence "ABCD EFG" would generate the following bigrams {"AB," "BC," "CD," "EF," and "FG"}. In our work, we generated these overlapping bigrams for Asian characters only, using Latin characters, digits, spaces and other punctuation marks (collected for each language in their respective encoding) to stop bigram generation. Moreover, we did not split any words written in ASCII characters. The most frequent terms may of course be removed before indexing. For the Chinese language, we defined a list of 39 most frequent unigrams, 49 most frequent bigrams and a list of 91 words (used when applying a word-based indexing scheme in Chinese). For Japanese we defined a short stopword list of 30 words and another of 20 bigrams, and for Korean our stoplist was composed of 91 bigrams.

Before generating the bigrams for the Japanese documents, we removed all Hiragana characters, given that these characters are used mainly to write words used only for grammatical purposes (e.g., *doing*, *in*, *of*), as well as inflectional endings for verbs, adjectives and nouns. Moreover, half-width characters were replaced by their corresponding full-width version.

For the English collection, we based the indexing process on the SMART stopword (571 words) and stemmer procedure.

1.3 Evaluation of various IR systems

To measure retrieval performance, we adopted non-interpolated mean average precision (MAP). To determine whether or not a given search strategy would be better than another, we based our statistical validation on the bootstrap approach [9]. In the tables appearing in this paper we have thus

underlined any statistically significant differences, with the means serving as baseline amounts (two-sided non-parametric bootstrap test, significance level at 5%).

MAP values obtained by the eleven search models under three query formulations (T, D, DN) are shown in Table 2 (for the English and Japanese collections), where the best performance for any given condition is shown in bold (these values were used as the baseline for our statistical tests in Tables 2 and 3). Table 3 lists performances obtained using the Korean (bigram) and Chinese (bigram or word) corpora.

Surprisingly, this data shows that the best retrieval scheme for short queries is not always the same as that for longer topics. For the Japanese collection (both bigram & word), the best retrieval models were always the PB2 when facing with short queries (T or D) and Okapi when using longer queries (DN). For the Chinese corpus (both bigram & word), the best retrieval model was always the PB2. Based on our statistical testing, the differences in performance were not always significant (e.g., for the Chinese corpus, the difference between Okapi and PB2 models is only significant for the D queries when using bigram indexing scheme). For the Japanese corpus, the word-based indexing scheme seemed to result in better retrieval performance. For example, based on the nine best performing IR models, and using T queries, the word-based indexing schemes shows, in average, a small 4.4% enhancement.

When comparing bigram and word-based representations for the Chinese collection (see Table 3), the performance difference seemed to favor more clearly word-based indexing. For example, based on the six best performing IR models and T queries, the average improvement was around 3.9% and favored the word-based IR schemes.

For the Korean corpus, increasing the query size from T to D improves, in average for the nine most effective IR models, the MAP of 8%, and of 28% for the DN over D query formulation.

1.4 Blind-query expansion

It was observed that pseudo-relevance feedback (blind-query expansion) seemed to be a useful technique for enhancing retrieval effectiveness. In this study, we adopted Rocchio's approach [2] whereby the system was allowed to add *m* terms extracted from the *k* best ranked documents from the original query using the following formula:

	Mean average precision								
	English (word, 49 queries)			Japanese (bigram, 47 queries)			Japanese (word, 47 queries)		
Model	T	D	DN	T	D	DN	T	D	DN
I(*n*)L2/PB2	0.3591	0.3548	**0.4556**	**0.2717**	**0.2829**	0.3957	**0.2895**	**0.3120**	0.3925
Okapi-npn	**0.3692**	0.3615	0.4555	0.2660	0.2694	**0.4079**	0.2655	0.2657	**0.4002**
Lnu-ltc	0.3562	0.3551	0.4185	0.2579	0.2648	0.3876	0.2743	0.2814	0.3780
dtu-dtn	0.3577	**0.3748**	0.3949	0.2461	0.2564	0.3660	0.2735	0.2944	0.3514
atn-ntc	0.3423	0.3458	0.3926	0.1799	0.1986	0.3287	0.2109	0.2335	0.3315
ltn-ntc	0.3275	0.3244	0.3608	0.2651	0.2538	0.3200	0.2723	0.2678	0.3115
ntc-ntc	0.2345	0.2522	0.3061	0.1292	0.1289	0.2302	0.1227	0.1343	0.1987
ltc-ltc	0.2509	0.2869	0.3675	0.0992	0.1104	0.2220	0.0945	0.1106	0.2106
lnc-ltc	0.2617	0.2868	0.3951	0.1070	0.1174	0.2354	0.1132	0.1236	0.2475
bnn-bnn	0.2000	0.1277	0.0964	0.1403	0.1422	0.1092	0.1403	0.0977	0.0564
nnn-nnn	0.1048	0.0701	0.0806	0.0981	0.0851	0.0900	0.1055	0.0477	0.0445

Table 2. MAP for various IR models (monolingual English and Japanese)

	Mean average precision								
	Korean (bigram, 50 queries)			Chinese (bigram, 50 queries)			Chinese (word, 50 queries)		
Model	T	D	DN	T	D	DN	T	D	DN
PB2	0.3729	**0.4141**	**0.5022**	**0.3042**	**0.2878**	**0.3973**	**0.3246**	**0.2974**	**0.4136**
Okapi-npn	0.3630	0.3823	0.4940	0.2995	0.2584	0.3887	0.3230	0.2816	0.4135
Lnu-ltc	**0.3973**	0.3962	0.4628	0.2999	0.2644	0.3667	0.3227	0.2910	0.3864
dtu-dtn	0.3673	0.3907	0.4497	0.2866	0.2565	0.3564	0.2894	0.2812	0.3760
atn-ntc	0.3270	0.3489	0.4541	0.2527	0.2378	0.3548	0.2578	0.2585	0.3668
ltn-ntc	0.3708	0.3688	0.4442	0.2886	0.2571	0.3421	0.2833	0.2570	0.3404
ntc-ntc	0.2506	0.2886	0.3666	0.2130	0.2093	0.3138	0.1645	0.1748	0.2741
ltc-ltc	0.2260	0.2638	0.3794	0.1933	0.2056	0.3382	0.1772	0.1931	0.3416
lnc-ltc	0.2414	0.2773	0.4172	0.2053	0.2115	0.3546	0.2189	0.2292	0.3754
bnn-bnn	0.2348	0.1840	0.1078	0.1629	0.1334	0.1139	0.1542	0.0915	0.0613
nnn-nnn	0.1770	0.1287	0.1911	0.1170	0.0911	0.1333	0.0738	0.0527	0.0468

Table 3. MAP for various IR models (monolingual Korean and Chinese)

$$Q' = \alpha \cdot Q + (\beta / k) \cdot \sum_{j=1}^{k} w_{ij} \quad (6)$$

in which Q' denotes the new query built for the previous query Q, and w_{ij} denotes the indexing term weight attached to the term t_j in the document D_i. In our evaluation, we fixed $\alpha = 0.75$, $\beta = 0.75$.

For a new blind-query expansion denoted IDFQE ("IDF Query Expansion"), we adopted the following procedure. First form the root set of search terms composed of all terms included in the original query Q and all indexing terms appearing in the *k* best ranked documents. The weight value for each term in this root set would be computed as follows:

$$w'_j = \alpha \cdot I_Q(tj) \cdot tf_j + (\beta / k) \cdot \sum_{i=1}^{k} I_{D_i}(t_j) \cdot idf_j \quad (7)$$

$$\text{with } I_Q(t_j) = 1 \text{ if } t_j \in Q,\ 0 \text{ otherwise}$$

where for term t_j, $idf_j = \ln(n/df_j)$ (the classical idf value) and $I_Q(t_j)$ (or $I_{Di}(t_j)$), an indicator function returning the value 1 if the term t_j belonging to the query Q (or the document D_i), otherwise the value is 0. In this weighting scheme, if a term appeared only in the original query Q, its weight would be $\alpha \cdot tf_j$, while a term appearing only in one document would have a weight of $(\beta / k) \cdot idf_j$.

The root set elements were then sorted in decreasing order according to their weight. To form the new query Q', we selected the top *m* search terms, and the weights attached to these selected terms in the new query were the same as those used in the root set. We thus used the same weighting scheme to select and weight the new search terms.

We used the two probabilistic models to evaluate this proposition. Table 4 summarizes some results achieved for the English, and Japanese (bigram and word-based indexing scheme) collections, while Table 5 shows some retrieval performances for the Korean (bigram) and Chinese (bigram or word-based indexing) corpora. In these tables, the rows labeled "PB2," (C, J, and K) "I(*n*)L2" (E) or "Okapi-npn" (baseline) indicate the MAP before applying this blind-query expansion procedure. The rows starting with "#doc/#term" indicate the number of top-ranked documents and the number of terms used to enlarge the original query. Finally, the rows labeled "& Rocchio" (or "& IDFQE") depict the MAP following Rocchio's approach (Eq. 6) (or our idf method, Eq. 7), and using the parameter setting specified in the previous row.

From the data shown in Tables 4 and 5, we could infer that the blind query expansion technique improved MAP, and this improvement was usually statistically significant (underlined values in these tables). When comparing both probabilistic models, this strategy seemed to perform better with the PB2 (or I(*n*)L2) than with the Okapi model. Moreover, enhancement percentages were greater for short topics than for longer ones. For example, in the Japanese

	Mean average precision								
	English (word, 49 queries)			Japanese (bigram, 47 queries)			Japanese (word, 47 queries)		
Model	T	D	DN	T	D	DN	T	D	DN
I(*n*)L2/PB2	0.3591	0.3548	0.4556	0.2717	0.2829	0.3957	0.2895	0.3120	0.3925
#doc/#term	15 / 100	15 / 100	10 / 60	15 / 100	10 / 75	10 / 100	15 / 100	20 / 120	10 / 80
& Rocchio	**0.4450**	**0.4625**	0.5027	0.3429	**0.3596**	**0.4240**	0.3479	0.3581	0.3983
#doc/#term	15 / 100	15 / 100	10 / 60	15 / 100	10 / 75	15 / 100	15 / 100	15 / 70	10 / 80
& IDFQE	0.4389	0.4543	**0.5039**	**0.3476**	0.3563	0.4180	**0.3690**	**0.3609**	**0.4071**
Okapi-npn	0.3692	0.3615	0.4555	0.2660	0.2694	0.4079	0.2655	0.2657	0.4002
#doc/#term	15 / 100	15 / 100	10 / 60	10 / 150	10 / 150	10 / 100	10 / 100	15 / 100	10 / 100
& Rocchio	0.4420	0.4478	0.4573	0.3266	0.3212	0.4103	0.3523	0.3433	0.4021
#doc/#term	15 / 100	15 / 100	10 / 60	15 / 100	15 / 100	15 / 100	20 / 100	20 / 100	10 / 100
& IDFQE	**0.4476**	**0.4529**	**0.4994**	**0.3501**	**0.3617**	**0.4307**	**0.3681**	**0.3763**	**0.4378**

Table 4. MAP with blind-query expansion (monolingual English and Japanese)

	Mean average precision								
	Korean (bigram, 50 queries)			Chinese (bigram, 50 queries)			Chinese (word, 50 queries)		
Model	T	D	DN	T	D	DN	T	D	DN
PB2	0.3729	0.4141	0.5022	0.3042	0.2878	0.3973	0.3246	0.2974	0.4136
#doc/#term	15 / 140	5 / 60	5 / 150	10 / 100	5 / 100	5 / 125	5 / 75	10 / 75	10 / 100
& Rocchio	0.3899	0.4719	0.5158	0.3782	0.3616	0.4241	0.3547	0.3822	0.4088
#doc/#term	15 / 100	10 / 100	15 / 100	10 / 75	10 / 125	5 / 125	5 / 125	10 / 75	10 / 100
& IDFQE	**0.4253**	**0.4766**	**0.5228**	**0.3912**	**0.3861**	**0.4288**	**0.3769**	**0.3954**	**0.4400**
Okapi-npn	0.3630	0.3823	0.4940	0.2995	0.2584	0.3887	0.3230	0.2816	0.4135
#doc/#term	15 / 100	5 / 100	15 / 200	5 / 125	10 / 100	5 / 125	5 / 75	10 / 75	10 / 100
& Rocchio	0.4346	0.4563	0.4881	**0.3559**	0.3176	0.3854	**0.3788**	0.3522	0.4252
#doc/#term	15 / 100	10 / 100	15 / 150	5 / 125	10 / 75	5 / 125	5 / 125	10 / 75	10 / 100
& IDFQE	**0.4453**	**0.4667**	**0.5304**	0.3557	**0.3659**	**0.4242**	0.3778	**0.3576**	**0.4479**

Table 5. MAP with blind query expansion (monolingual Korean and Chinese)

	Mean average precision							
	English (word, 49 queries)			Japanese (word or bigam, 47 queries)				
Model	T	D	DN	T	T	D	D	DN
#doc/#term	I(*n*)L2 (wd) 15 / 50 R *0.4425*	I(*n*)L2 (wd) 20 / 70 R 0.4494	PB2 (wd) 15 / 40 I 0.4589	Okapi (wd) 20 / 100 I 0.3681	PB2 (wd) 15 / 100 I 0.3690	Okapi (wd) 20 / 100 I 0.3763	Okapi (wd) 20 / 100 I 0.3763	Okapi (wd) 10 / 100 I 0.4378
#doc/#term	Okapi (wd) 15 / 100 R 0.4420	Okapi (wd) 15 / 100 R *0.4478*	I(*n*)L2 (wd) 10 / 60 R 0.5027	Okapi (bi) 15 / 100 I 0.3501	Okapi (wd) 10 / 100 R 0.3523	Okapi (bi) 15 / 100 I 0.3617	Okapi (wd) 15 / 100 R 0.3433	Okapi (bi) 15 / 150 I 0.4307
Round-rob.	0.4427	0.4514	0.4942	0.3639	0.3729	0.3761	0.3708	0.4405
SumRSV	**0.4544**	0.4573	0.5018	0.3637	0.3742	0.3752	**0.3742**	0.4486
NormRSV	*0.4539*	0.4575	*0.5019*	*0.3734*	*0.3839*	0.3780	0.3681	0.4496
Z-score	0.4540	***0.4581***	**0.5039**	0.3693	**0.3852**	*0.3773*	*0.3692*	***0.4504***
Z-score W	0.4517	0.4572	0.4982	**0.3754**	0.3839	**0.3801**	0.3736	0.4499

Table 6. MAP with various data fusion schemes (English and Japanese corpora)

collection (word-based indexing) using the PB2 model and T topics, blind query expansion improved mean performance, ranging from 0.2895 to 0.3479 (+20.1% in relative effectiveness) with Rocchio's approach or to 0.3690 with IDFQE (+27.5%). With DN query formulation, the MAP improves from 0.3925 to 0.4071 (+3.7%) using our IDFQE scheme.

1.5 Data fusion

For a strategy that would enhance retrieval effectiveness, we can combine two or more result lists. As a first data fusion strategy, we considered the round-robin approach whereby we selected one document in turn from all individual lists and removed duplicates, retaining the highest ranking instances. Various other data fusion operators were suggested [4], however the simple linear combination (denoted "SumRSV") usually seemed to provide the best performance [10], [4], or at least good overall performance [11]. For a given set of result lists i = 1, 2, …, r, this combined operator was defined as $\text{SumRSV} = \sum \text{RSV}_i$, being the simple sum of all document scores (RSV_i) obtained by each search model.

As a third data fusion strategy we normalized document scores within each collection through dividing them by the maximum score. As a variant of this normalized score merging scheme (denoted "NormRSV"), we might normalize the document RSV_k scores within the *i*th result list, as follows:

$$\text{NormRSV}_k = ((\text{RSV}_k - \text{Min}^i) / (\text{Max}^i - \text{Min}^i)) \quad (8)$$

As a fourth data fusion strategy, we suggest merging the retrieved documents according to the Z-score, computed for each result list. Within this scheme, we would normalize retrieval status values for each document D_k provided by the *i*th result list, as computed by the following formula:

$$\text{Z-score RSV}_k = \gamma_i \cdot [((\text{RSV}_k - \text{Mean}^i) / \text{Stdev}^i) + \delta^i],$$
$$\delta^i = ((\text{Mean}^i - \text{Min}^i) / \text{Stdev}^i) \quad (9)$$

within which Mean^i denotes the average of the RSV_k, Stdev^i the standard deviation, and γ_i (usually fixed at 1), used to reflect the retrieval performance of the underlying retrieval model. When the coefficients γ_i are not all fixed at 1, the data fusion operator is denoted as "Z-score W."

	Mean average precision								
	Chinese (word, bigram, unibigram, 50 queries)					Korean (bigram, 50 queries)			
Model	T	T	D	D	DN	T	T	D	DN
#doc/#term	PB2 (wd) 5 / 75 R 0.3547	PB2 (wd) 5 / 75 R 0.3547	PB2 (wd) 10 / 75 I 0.3954	PB2 (wd) 10 / 75 R 0.3822	PB2 (wd) 10 / 100 I 0.4400	Okapi (bi) 15 / 100 I 0.4453	Okapi (bi) 15 / 100 I 0.4453	Okapi (bi) 5 / 100 R 0.4563	Okapi (bi) 15 / 150 I 0.5304
#doc/#term	PB2 (bi) 10 / 75 I 0.3912	Okapi (wd) 5 / 125 I 0.3778	Okapi (wd) 10 / 75 I 0.3576	Okapi (wd) 10 / 75 I 0.3576		PB2 (bi) 15 / 100 I 0.4253	Okapi (bi) 15 / 100 R 0.4346	PB2 (bi) 10 / 100 I *0.4766*	PB2 (bi) 5 / 150 R 0.5158
#doc/#term	Oka(unibi) 5 / 100 I 0.3620		PB2(unibi) 10 / 100 I 0.3738	PB2(bi) 5 / 100 R 0.3616	PB2(unibi) 10 / 100 I 0.4557				
Round-rob.	0.3780	0.3691	0.3850	0.3734	0.4498	0.4393	0.4463	0.4746	0.5267
SumRSV	**0.4121**	0.3712	*0.4057*	0.3956	0.4618	0.4351	0.4526	0.4892	0.5293
NormRSV	0.4062	0.3800	0.4064	0.4006	0.4592	0.4396	*0.4525*	0.4913	0.5333
Z-score	*0.4076*	*0.3828*	0.4091	***0.4026***	*0.4585*	*0.4395*	**0.4547**	***0.4921***	*0.5362*
Z-score W	0.4050	**0.3837**	**0.4127**	0.3980	**0.4593**	**0.4415**	0.4545	0.4900	**0.5383**

Table 7. MAP with various data fusion schemes (Chinese and Korean corpora)

We could of course combine different document surrogates during the indexing process. For the Chinese corpus for example, we might index the documents (and the queries) using both unigram (or character) and bigram approaches (denoted by the label "unibi" in this paper).

Table 6 shows the MAP obtained for the English and Japanese collections, for each of the T, D and DN queries. Table 7 lists the same information for the Chinese and Korean corpora, in which the best performing single IR scheme served as a baseline for our statistical testing.

From this data, we could see that combining two or more IR models might sometimes improve retrieval effectiveness (differences with the best single system were however not statistically significant except three cases with the Japanese corpus). Moreover the Z-score scheme tended to produce the best performance. It is difficult however to predict which data fusion operator would produce the best result, even when a particular data fusion scheme improved performance during single runs. Current and past experiments tend to indicate that combining short query results provides more improvement than does combining longer topics [11].

Results from some of our official monolingual runs are indicated in italics in shown in Tables 6 and 7. Given that we introduced a bug in our IDFQE blind-query expansion scheme, our official results depicted usually a lower MAP than the corrected version (differences given in Table 8).

	Official MAP	Corrected MAP
UniNE-J-J-DN-01	**0.4480**	0.4504
UniNE-J-J-T-02	0.3705	0.3734
UniNE-J-J-D-03	**0.3823**	0.3773
UniNE-J-J-T-04	0.3815	0.3852
UniNE-J-J-D-05	0.3717	0.3692
UniNE-C-C-DN-01	**0.4419**	0.4585
UniNE-C-C-T-02	0.4104	0.4076
UniNE-C-C-D-03	0.3846	0.4057
UniNE-C-C-T-04	0.3806	0.3828
UniNE-C-C-D-05	0.4002	0.4026
UniNE-K-K-DN-01	0.5313	0.5362
UniNE-K-K-T-02	0.4494	0.4395
UniNE-K-K-D-03	0.4845	0.4921
UniNE-K-K-T-04	0.4468	0.4525
UniNE-K-K-D-05	0.4748	0.4766

Table 8. Official and corrected results

For the Japanese monolingual task for example, the UniNE-J-J-T-02 was based on the "NormRSV" operator combining two Okapi runs (the first was a word-based indexing scheme and the second based on bigrams) with an official MAP 0.3705 and a corrected MAP of 0.3734.

2 Bilingual IR

As explained in our last NTCIR campaign paper [11], we translated each topic written in English into the three Asian languages using freely available resources on the Web. In this study, we chose four different machine translation (MT) systems and three machine-readable bilingual dictionaries (MRDs) to translate the topics:

SYSTRAN	www.systranlinks.com
WORLDLINGO	www.worldlingo.com
ALPHAWORKS	www.alphaWorks.ibm.com
APPLIEDLANGUAGE	www.appliedLanguage.com
DICT	www.dicts.info
ECTACO	www.ectaco.co.uk/free-online-dictionaries
BABYLON	www.babylon.com

For the bilingual dictionaries, we submitted search keywords word-by-word after lemmatizing (e.g., "weapons" will be replaced by "weapon"). In response to each word submitted, the MRD system provided not only one but several translation terms (in an unspecified order). In our experiments, we decided to pick the first available translation (e.g., labeled "Babylon 1" or "Dict 1"), the first two (e.g., "Babylon 2") or the first three (e.g., "Dict 3").

Table 9 shows MAP when translating English topics employing the four MT systems, the three MRDs and the Okapi model. This table also contains the retrieval performance for manually translated topics, with the first row ("Okapi-npn") being used as a baseline. Compared to our previous work with European languages [10] and also to manually translated topics, machine translated topics generally provided poor performance levels. Based on the T queries and the best single query translation resource (the Alphawork MT system in this case), the resulting performance was only 40.3% that of a monolingual search for the Chinese language (0.1208 vs. 0.2995), 56.6% for the Korean language (0.2055 vs. 0.3630) or 69.7% for the Japanese (0.1855 vs. 0.2660). Moreover, differences in mean average precision were always statistically significant and favored the manual topic translation approach.

The Alphawork MT system seemed to produce the best translated topics for all languages. Morevoer, MT systems tended to result in better performance level than MRDs approaches. The poor overall query translation performance seemed to be caused by including proper names in numerous topics (e.g, 15 topics had a person's name, 4 a geographical name, 7 had other proper names such as "Linux," "Anthrax" or "Mir"), and these names were usually not properly translated by the MRDs or MT systems.

3 Multilingual IR

In this section, we will investigate situations in which users submit a topic in English in order to retrieve relevant documents in English, Chinese, Japanese and Korean (CJKE). The different collections were indexed separately and, once the original or translated request (see Section 2) was received, a ranked list of retrieved items was returned. From these lists we needed to produce a unique ranked result list, using a merging strategy described further on in this section.

	Mean average precision								
	Chinese (bigram, 50 queries)			Japanese (bigram, 47 queries)			Korean (bigram, 50 queries)		
Model	T	D	DN	T	D	DN	T	D	DN
Okapi-npn	0.2995	0.2584	0.3887	0.2660	0.2694	0.4079	0.3630	0.3823	0.4940
Babylon 1	0.0505	0.0486	0.1059	0.0987	0.1161	0.1467	n/a	n/a	n/a
Babylon 2	0.0433	0.0516	0.0943	0.1250	0.1137	0.1375	n/a	n/a	n/a
Babylon 3	0.0438	0.0480	0.1113	0.1191	0.1212	0.1329	n/a	n/a	n/a
Ectaco 1	n/a	n/a	n/a	n/a	n/a	n/a	0.0632	0.0392	0.0500
Dict 1	0.0411	0.0329	0.0249	0.0570	0.0248	0.0366	0.0473	0.0373	0.0287
Dict 2	0.0700	0.0495	0.0552	0.0736	0.0341	0.0411	0.0644	0.0715	0.0615
Dict 3	0.0715	0.0540	0.0630	0.0745	0.0314	0.0407	0.0780	0.0767	0.0781
WorldLing	0.1055	0.1252	0.2256	0.1417	0.1597	0.2637	0.1988	0.2113	0.3418
AlphaW	**0.1208**	**0.1663**	**0.2526**	**0.1855**	**0.2021**	**0.3037**	**0.2055**	0.2117	0.3363
AppliedLg	0.1052	0.1255	0.2269	0.1417	0.1609	0.2642	0.1988	**0.2118**	**0.3421**
Systran	0.1052	0.1255	0.2269	0.1417	0.1609	0.2642	0.1988	0.2113	0.3415
Combined	indexing : bigram only			Systran / WorldLingo / AlphaWorks					
with Okapi	0.1317	0.1689	0.2713	**0.1927**	0.2039	**0.3056**	0.2396	0.2557	0.3914
with PB2	0.1355	**0.1946**	**0.2816**	0.1925	0.2214	0.2937	0.2503	0.2848	0.4060

Table 9. MAP for various query translation approaches (Okapi model)

	Mean average precision				
	T	T	D	D	DN
English	Oka& DFR	Okapi	Oka& DFR	DFR	Oka& DFR
(out of 49 queries)	0.4540	0.4420	0.4572	0.4494	0.5019
Chinese	Oka& PB2	Okapi(wd)	PB2 & PB2	PB2 (wd)	PB2 & PB2
(out of 50 queries)	0.2417	0.2360	0.2751	0.2363	0.2904
Japanese	Oka& Oka	Okapi(wd)	Oka& Oka	Okapi(wd)	Oka& Oka
(out of 47 queries)	0.2631	0.2631	0.2878	0.2728	0.3379
Korean	Oka& PB2	Okapi(bi)	Oka& DFR	PB2 (bi)	Oka& PB2
(out of 50 queries)	0.3374	0.3289	0.3586	0.3677	0.4120
Merging strategy CJKE					
Round-robin (baseline)	0.2244	0.2169	0.2548	0.2410	0.2839
Raw-score	0.2165	**0.2332**	0.2364	0.2169	0.2823
MaxRSV	0.2248	0.2102	0.2468	0.1979	0.2830
NormRSV (Eq. 8)	0.2256	0.2259	*0.2475*	0.2322	0.2830
Biased RR E,K=1/C,J=2	0.2036	0.1965	0.2328	0.2172	0.2600
Z-score (Eq. 9)	**0.2333**	*0.2261*	**0.2698**	***0.2578***	**0.2950**
Z-score W E,K=1/C,J=1.25	*0.2113*	0.1965	0.2475	0.2316	*0.2695*

Table 10. MAP of various merging strategies for CJKE collection (official in italics)

As a first approach, we considered the round-robin method. As a second merging approach, we took the document score into account, denoted as RSV_k for document D_k. Known as raw-score merging, this strategy, produced a final list sorted by document score, as computed by each collection. As a third scheme, we could either normalize the RSV_k by using the document score of the retrieved record in the first position ("MaxRSV") or using Eq. 8 ("NormRSV").

As a fifth merging scheme, we would suggest a biased round-robin approach which extracted not just one document per collection per round, but one document from both the English and Korean collections and two from the Japanese and Chinese. This type of merging strategy exploited the fact that the Japanese and Chinese corpora contain more articles than do the English or the Korean corpora. Finally, we applied our Z-score (see Eq. 9) and then under the "Z-score W" label we assigned a weight of 1.25 for the Japanese and Chinese result lists, and 1 for the English and Korean runs.

The data depicted in Table 10 also indicates that resulted in retrieval effectiveness that could be viewed as statistically superior to that of the round-robin baseline. As a first approach, the simple and normalized merging schemes ("MaxRSV" or "NormRSV") provided reasonable performance levels. Also, our biased round-robin scheme did not perform better when compared to the simple round-robin version (it was difficult a priori to know whether a given corpus would really contain more relevant items than another). The Z-score provided statistically better performance levels than did the round-robin approach.

Conclusion

Based on our evaluations, we may infer that for both the Chinese or Japanese language, using a good automatic word-segmentation procedure seems to produce slightly better retrieval performances than an bigram indexing scheme (average difference between 3.9% and 7%, see Tables 2 and 3). Based on our evaluation of the various IR models, we can obtain the best retrieval performance levels using the PB2 probabilistc model before blind-query expansion, and using Okapi after blind-query expansion for the Japanese and Korean languages (Tables 2 through 5).

Compared to Rocchio's query expansion (Eq. 6), better performance may be obtained from our idf-based model (see Eq. 7). The performance differences with an approach without query expansion are usually statistically significant and in favor of a query expanded approach. To further improve retrieval effectiveness, a data fusion approach could also be considered, although this technique would require additional computational resources with and uncertain improvement (Tables 6 and 7).

From an analysis of bilingual search performances, the number and quality of freely available translation resources were questionable. When translating the topics from English into Chinese, Japanese or Korean language, overall retrieval effectiveness decreases by more than 30% for the Japanese language, compared to more than 50% for Chinese and Korean (see Table 9).

When evaluating various merging strategies (see Table 10), it appears that the Z-score merging procedure produces better retrieval performance when result lists provided by separate collections are merged.

Acknowledgments

This research was supported in part by the Swiss NSF (Grant #200020-103420).

References

[1] Amati, G., & van Rijsbergen, C.J. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM-TOIS*, 20(4):357-389, 2002.

[2] Buckley, C., Singhal, A., Mitra, M., & Salton, G. New retrieval approaches using SMART. *Proceedings of TREC-4*, pp. 25-48, 1996.

[3] Chen, A., & Gey, F.C. Experiments on cross-language and patent retrieval at NTCIR-3 workshop. *Proceedings of NTCIR-3, 2003*.

[4] Fox, E.A., & Shaw, J.A. Combination of multiple searches. *Proceedings TREC-2*, pp. 243-249, 1994.

[5] Kishida, K., Chen, K.-H., Lee, S., Kuriyama, K., Kando, N., Chen, H.-H., & Myaeng, S.H. Overview of CLIR Task at the Fifth NTCIR Workshop. *Proceedings of NTCIR-5*, Tokyo, 2005.

[6] Luk, R.W.P., & Kwok, K.L.. A comparison of Chinese document indexing strategies and retrieval models. *ACM-TALIP*, 1(3): 225-268, 2002.

[7] Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., & Asahara, M. Japanese morphological analysis system ChaSen. Technical Report NAIST-IS-TR99009, NAIST, 1999 (available at http://chasen.aist-nara.ac.jp/).

[8] Robertson, S.E., Walker, S., & Beaulieu, M. Experimentation as a way of life: Okapi at TREC. *IP&M*, 36(1), 95-108, 2000.

[9] Savoy, J. Statistical inference in retrieval effectiveness evaluation. *IP&M*, 33(4):495-512, 1997.

[10] Savoy, J. Combining multiple strategies for effective monolingual and cross-lingual retrieval. *IR Journal*, 7(1-2):121-148, 2004.

[11] Savoy, J. Comparative study of monolingual and multilingual search models for use with Asian languages. ACM TALIP, 4(3), 2005.

[12] Singhal, A., Choi, J., Hindle, D., Lewis, D.D., & Pereira, F. AT&T at TREC-7. *Proceedings of TREC-7*, 239-251, 1999.

Appendix

bnn	$w_{ij} = 1$	npn	$w_{ij} = tf_{ij} \cdot \ln[(n-df_j) / df_j]$	ltn	$w_{ij} = [\ln(tf_{ij}) + 1] \cdot idf_j$
nnn	$w_{ij} = tf_{ij}$	lnc	$w_{ij} = \dfrac{\ln(tf_{ij}) + 1}{\sqrt{\sum_{k=1}^{t} (\ln(tf_{ik}) + 1)^2}}$	ntc	$w_{ij} = \dfrac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t} (tf_{ik} \cdot idf_k)^2}}$
atn	$w_{ij} = idf_i \cdot [0.5+0.5 \cdot tf_{ij}/maxtf_i]$			dtn	$w_{ij} = [\ln[\ln(tf_{ij})+1] + 1] \cdot idf_j$
ntn	$w_{ij} = tf_{ij} \cdot idf_j$				
Lnu	$w_{ij} = \dfrac{\left(1 + \ln(tf_{ij})\right) / \left(\ln(\text{mean tf}) + 1\right)}{(1 - \text{slope}) \cdot \text{pivot} + \text{slope} \cdot nt_i}$			ltc	$w_{ij} = \dfrac{\left(\ln(tf_{ij}) + 1\right) \cdot idf_j}{\sqrt{\sum_{k=1}^{t} \left(\left(\ln(tf_{ik}) + 1\right) \cdot idf_k\right)^2}}$
Okapi	$w_{ij} = \left((k_1 + 1) \cdot tf_{ij}\right) / \left(K + tf_{ij}\right)$			dtu	$w_{ij} = \dfrac{\left(\ln\left(\ln(tf_{ij}) + 1\right) + 1\right) \cdot idf_j}{(1 - \text{slope}) \cdot \text{pivot} + \left(\text{slope} \cdot nt_i\right)}$

Table A.1. Weighting schemes

Statistical and Comparative Evaluation of Various Indexing and Search Models

Samir Abdou and Jacques Savoy

Computer Science Department, University of Neuchatel,
rue Emile Argand 11, 2009 Neuchatel, Switzerland
{Samir.Abdou, Jacques.Savoy}@unine.ch

Abstract. This paper first describes various strategies (character, bigram, automatic segmentation) used to index the Chinese (ZH), Japanese (JA) and Korean (KR) languages. Second, based on the NTCIR-5 test-collections, it evaluates various retrieval models, varying from classical vector-space models to more recent developments in probabilistic and language models. While no clear conclusion was reached for the Japanese language, the bigram-based indexing strategy seems to be the best choice for Korean, and the combined "unigram & bigram" indexing strategy is best for traditional Chinese. On the other hand, *Divergence from Randomness* (DFR) probabilistic model usually results in the best mean average precision. Finally, upon an evaluation of the four different statistical tests, we find that their conclusions correlate, even more when comparing the non-parametric bootstrap with the t-test.

1 Introduction

In order to promote IR activities involving Asian languages and also to facilitate technological transfers into products, the latest NTCIR evaluation campaign [1] created test-collections for the traditional Chinese, Japanese and Korean languages. Given that English is an important language for Asia and that we also wanted to verify that the various approaches suggested might also work well with European languages, a fourth collection of newspaper articles written in English was used.

Even with all participants working with the same newspapers corpora and queries, it is not always instructive to directly compare IR performance results achieved by two search systems. In fact, given that their performance is usually based on different indexing and search strategies involving a large number of underlying variables (size and type of stopword lists, stemming strategies, token segmentation, n-grams generation procedures, indexing restrictions or adaptations and term weighting approaches).

Based on the NTCIR-5 test-collections [1], this paper empirically compares various indexing and search strategies involving East Asian languages. In order to obtain more solid conclusions, this paper also considers various IR schemes, and all comparisons are analyzed statistically. The rest of this paper is organized as follows: Section 2 describes the main features of the test-collections.

H.T. Ng et al. (Eds.): AIRS 2006, LNCS 4182, pp. 362–373, 2006.

Section 3 contains an overview of the various search models, from vector-space approaches to recent developments in both probabilistic and language models. Section 4 portrays the different indexing strategies used to process East Asian languages, and Section 5 contains various evaluations and analyzes of the resultant retrieval performance. Finally, Section 6 compares decisions that might result from using other statistical tests and Section 7 presents the main findings of our investigation.

2 Overview of NTCIR-5 Test-Collections

The test-collections used in our experiments include various newspapers covering the years 2000-2001 [1]. The Chinese and Japanese corpora were larger in size (1,100 MB) but the Chinese collection contained a slightly larger number of documents (901,446) than did the Japanese (858,400). The Korean and English corpora were smaller, both in terms of size (438 MB for the English and 312 MB for the Korean) and number of newspaper articles (259,050 for the English and 220,374 for the Korean).

When analyzing the number of pertinent documents per topic, only rigid assessments were considered, meaning that only "highly relevant" and "relevant" items were viewed as being relevant, under the assumption that only highly or relevant items would be useful for all topics. A comparison of the number of relevant documents per topic indicates that for the English collection the median number of relevant items per topic is 33, while for the Asian languages corpora it is around 25 (ZH: 26, JA: 24, KR: 25.5). The number of relevant articles is also greater for the English (3,073) corpus, when compared to the Japanese (2,112), Chinese (1,885) or Korean (1,829) corpora.

The 50 available topics covered various subjects (e.g., "Kim Dae-Jun, Kim Jong Il, Inter-Korea Summit," or "Harry Potter, circulation"), including both regional/national events ("Mori Cabinet, support percentage, Ehime-maru") or topics having a more international coverage ("G8 Okinawa Summit"). The same set of queries was available for the four languages, namely Chinese, Japanese, Korean and English. According to the TREC model, the structure of each topic consisted of four logical sections: brief title (<TITLE>), one-sentence description (<DESC>), narrative (<NARR>) specifying both the background context (<BACK>) and a relevance assessment criterion (<REL>) for the topic. Finally a concept section (<CONC>) provides some related terms. In our experiments, we only use the title field of the topic description.

3 Search Models

In order to obtain a broader view of the relative merit of the various retrieval models, we examined six vector-space schemes and three probabilistic models. First we adopted the classical $tf\ idf$ model, in which the weight (denoted w_{ij}) attached to each indexing term t_j in document D_i was the product of its term occurrence frequency (or tf_{ij}) and its inverse document frequency (or

$idf_j = ln(n/df_j)$, where n indicates the number of documents in the corpus, and df_j the number of documents in which the term t_j appears). To measure similarities between documents and requests, we computed the inner product after normalizing indexing weights (model denoted "document=ntc, query=ntc" or "ntc-ntc").

Other variants might also be created, especially in cases when the occurrence of a particular term in a document is considered as a rare event. Thus, the proper practice may be to give more importance to the first occurrence of a term, as compared to any successive occurrences. Therefore, the tf component might be computed as the $ln(tf) + 1$ (denoted "ltc", "lnc", or "ltn") or as $0.5 + 0.5 \cdot [tf \ / \ max \ tf \ in \ D_i]$ ("atn"). We might also consider that a term's presence in a shorter document would be stronger evidence than its occurrence in a longer document. More complex IR models have been suggested to account for document length, including the "Lnu" [2], or the "dtu" IR models [3] (more details are given in the Appendix).

In addition to vector-space approaches, we also considered probabilistic IR models, such as the Okapi probabilistic model (or BM25) [4]. As a second probabilistic approach, we implemented the PB2 taken from the *Divergence from Randomness* (DFR) framework [5], based on combining the two information measures formulated below:

$w_{ij} = Inf^1_{ij}(tf) \cdot Inf^2_{ij}(tf) = -log_2 \left[Prob^1_{ij}(tf)\right] \cdot (1 - Prob^2_{ij}(tf))$

where w_{ij} indicates the indexing weight attached to term t_j in document D_i, $Prob^1_{ij}(tf)$ is the pure chance probability of finding tf_{ij} occurrences of the indexing unit t_j in the document D_i. On the other hand, $Prob^2_{ij}(tf)$ is the probability of encountering a new occurrence of t_j in the document given that we have already found tf_{ij} occurrences of this indexing unit. Within this framework, the PB2 model is based on the following formulae:

$$Prob^1_{ij}(tf) = \left[e^{\lambda_j} \cdot \lambda_j^{tfn_{ij}}\right] / \ tf_{ij}! \quad \text{with } \lambda_j = tc_j/n \tag{1}$$

$$Prob^2_{ij}(tf) = 1 - \left[\frac{tc_j + 1}{df_j \cdot (tfn_{ij} + 1)}\right] \quad \text{with} \tag{2}$$

$$tfn_{ij} = tf_{ij} \cdot \log_2 [1 + ((c \cdot mean \ dl)/l_i)] \tag{3}$$

where tc_j indicates the number of occurrences of t_j in the collection, *mean dl* the mean length of a document and l_i the length of document D_i.

Finally, we also considered an approach based on a language model (LM) [6], known as a non-parametric probabilistic model (the Okapi and PB2 are viewed as parametric models). Probability estimates would thus not be based on any known distribution (as in Equation 1) but rather be estimated directly, based on occurrence frequencies in document D or corpus C. Within this language model paradigm, various implementations and smoothing methods might also be considered, and in this study we adopted a model proposed by Hiemstra [6], as described in Equation 4, which combines an estimate based on document ($P[t_j \mid D_i]$) and corpus ($P[t_j \mid C]$).

$$P[Di|\ Q] \ = \ P[D_i] \ \cdot \ \prod_{t_j \in Q} [\lambda_j \cdot P[t_j \mid D_i] + (1 - \lambda_j) \cdot P[t_j \mid C]] \qquad (4)$$

with $P[t_j \mid D_i] = tf_{ij}/l_i$, $P[t_j \mid C] = df_j/lc$, $lc = \sum_k df_k$, and where λ_j is a smoothing factor (fixed at 0.3 for all indexing terms t_j) and lc an estimate of the corpus size.

4 Indexing Strategies

In the previous section, we described how each indexing unit was weighted to reflect its relative importance in describing the semantic content of a document or a request. This section will explain how such indexing units are extracted from documents and topic formulations.

For the English collection, we used words as indexing units and we based the indexing process on the SMART stopword list (571 terms) and stemmer. For European languages, it seems natural to consider words as indexing units, and this assumption has been generally confirmed by previous CLEF evaluation campaigns [7].

For documents written in the Chinese and Japanese languages, words are not clearly delimited. We therefore indexed East Asian languages using an overlapping bigram approach, an indexing scheme found to be effective for various Chinese collections [8], [9]. In this case, the "ABCD EFG" sequence would generate the follow-ing bigrams "AB," "BC," "CD," "EF," and "FG". Our choice of an indexing tool also involves other factors. As an example for Korean, Lee *et al.* [10] found more than 80% of nouns were composed of one or two Hangul characters, while for Chinese Sproat [11] reported a similar finding. An analysis of the Japanese corpus reveals that the mean length of continuous Kanji characters to be 2.3, with more than 70% of continuous Kanji sequences being composed of one or two characters (for Hiragana: mean=2.1, for Katakana: mean=3.96).

In order to stop bigram generation in our work, we generated overlapping bigrams for Asian characters only, using spaces and other punctuation marks (as collected for each language from its respective encoding). Moreover, in our experiments, we did not split any words written in ASCII characters, and the most frequent bigrams were removed before indexing. As an example, for the Chinese language we defined and removed a list of 90 most frequent unigrams, 49 most frequent bigrams and 91 most frequent words. For the Japanese language, we defined a stopword list of 30 words and another of 20 bigrams, and for Korean our stoplist was composed of 91 bigrams and 85 words. Finally, as suggested by Fujii & Croft [12], before generating bigrams for the Japanese documents we removed all Hiragana characters, given that these characters are mainly used to express grammatical words (e.g., *doing*, *do*, *in*, *of*), and the inflectional endings of verbs, adjectives and nouns. Such removal is not error-free because Hiragana could also be used to write Japanese nouns.

For Asian languages, there are of course other indexing strategies that might be used. In this vein, various authors have suggested that words generated by a segmentation procedure could be used to index Chinese documents. Nie &

Ren [13] however indicated that retrieval performance based on word indexing does not really depend on an accurate word segmentation procedure and this was confirmed by Foo & Li [14]. They also stated that segmenting a Chinese sentence does affect retrieval performance and that recognizing a greater number of 2-characters words usually contributes to retrieval enhancement. These authors did not however find a direct relationship between segmentation accuracy and retrieval effectiveness. Moreover, manual segmentation does not always result in better performance when compared to character-based segmentation.

To analyze these questions, we also considered automatic segmentation tools, namely Mandarin Tools (MTool, `www.mandarintools.com`) for the traditional Chinese language and the Chasen (`chasen.aist-nara.ac.jp`) morphological analyzer for Japanese. For Korean, the presence of compound construction could harm retrieval performance. Thus, in order to automatically decompose them, we applied the Hangul Analyser Module (HAM, `nlp.kookmin.ac.kr`) tool. With this linguistic approach, Murata *et al.* [15] obtained effective retrieval results while Lee *et al.* [9] showed that n-gram indexing could result in similar and sometimes better retrieval effectiveness, compared to word-based indexing applied in conjunction with a decompounding scheme.

5 Evaluation of Various IR Models

To measure retrieval performance, we adopted mean average precision (MAP) as computed by `TREC_EVAL`. To determine whether or not a search strategy might be better than another, we applied a statistical test. More precisely, we stated the null hypothesis (denoted H_0) specifying that both retrieval schemes achieved similar performance levels (MAP), and this hypothesis would be rejected at the significance level fixed at $\alpha = 5\%$ (two-tailed test). As a statistical test, we chose the non-parametric bootstrap test [16]. All evaluations in this paper were based on the title-only query formulation.

The MAP achieved by the six vector-space schemes, two probabilistic approaches and the language model (LM) are shown in Table 1 for the English and Chinese collections. The best performance in any given column is shown in bold and this value served as baseline for our first set of statistical tests. In this case, we wanted to verify whether this highest performance was statistically better than other performances depicted in the same column. When performance differences were detected as significant, we placed an asterisk (*) next to a given search engine performance. In the English corpus for example, the PB2 model achieved the highest MAP (0.3728). The difference in performance between this model and the "Lnu-ltc" approach (0.3562) was statistically significant while the difference between it and the Okapi model (0.3692) was not significant.

For the Chinese corpus, the PB2 probabilistic model also resulted in the best performance, except for the unigram-based indexing scheme where the best performance was obtained by the language model LM (0.2965). With these various indexing schemes, the difference between either the PB2, the LM, the Okapi or the "Lnu-ltc" models were not statistically significant. PB2 was the

Table 1. MAP for English and Chinese corpora (T queries)

	Mean average precision (MAP)				
	English	Chinese			
Model	word	unigram	bigram (base)	MTool	uni+bigram
PB2-nnn	**0.3728**	0.2774	**0.3042**	**0.3246**	**0.3433**
LM	0.3428*	**0.2965**	0.2594*	0.2800*	0.2943*
Okapi-npn	0.3692	0.2879	0.2995	0.3231	0.3321
Lnu-ltc	0.3562*	0.2883	0.2999	0.3227	0.3356
dtu-dtn	0.3577	0.2743	0.2866	0.2894*	0.3094*
atn-ntc	0.3423*	0.2329*	0.2527*	0.2578*	0.2729*
ltn-ntc	0.3275*	0.2348*	0.2886	0.2833*	0.3068*
ltc-ltc	0.2509*	0.1464	0.1933*	0.1772*	0.2202*
ntc-ntc	0.2345*	0.1162*	0.2130*	0.1645*	0.2201*
Improvement (7 best mod.)		-5.0%	0%	+4.5%	+10.2%

preferred model but by slightly changing the topic set, other models might perform better.

Based on an analysis of the four different indexing schemes used with the Chinese corpus, the data in Table 1 indicates that the combined "uni+bigram" indexing scheme tends to result in the best performance levels. As shown in the last row of this table, we computed mean improvements over the bigram indexing strategy, considering only the 7-best performing IR models (rows ending with the "ltn-ntc" model). From this overall measure we can see for example that the character-based indexing strategy results in lower performance level than does the bigram scheme (-5.0%). Using the bigram indexing strategy as a baseline, we verified whether performance differences between the various indexing schemes were statistically significant, and then underlined those that were statistically significant. Table 1 illustrates that the differences between the bigram and word-based indexing strategies (row labeled "MTool") are usually not significant. The differences between the bigram approach and the combined indexing strategy (last column) are usually significant and in favor of the combined approach.

Table 2. MAP for Japanese corpus (T queries)

	Mean average precision (MAP)			
Model	unigram	bigram (base)	Chasen	uni+bigram
PB2-nnn	**0.2240**	**0.2816**	**0.3063**	**0.3026**
LM	0.1369*	0.1791*	0.1968*	0.1944*
Okapi-npn	0.2208	0.2660*	0.2655*	0.2802
Lnu-ltc	0.2239	0.2579*	0.2743*	0.2736
dtu-dtn	0.2126	0.2461*	0.2735*	0.2735
atn-ntc	0.1372*	0.1799*	0.2109*	0.1901*
ltn-ntc	0.1518*	0.2651	0.2723	0.2726*
ltc-ltc	0.0580*	0.0992*	0.0945*	0.1154*
ntc-ntc	0.0706*	0.1292*	0.1227*	0.1295*
Improvement	-22.0%	0%	+7.4%	+6.6%

Evaluations done on the Japanese corpus are given in Table 2. With this language, the best performing search model was always PB2, often showing significant improvement over others (indicated by "*"). Comparing the differences between the four indexing strategies shows that both Chasen (automatic segmentation) and the combined indexing approaches ("uni+bigram") tend to result in the best performance levels. Using the bigram indexing strategy as baseline, the differences between the word (Chasen) or the combined ("uni+bigram") indexing strategies are however usually not significant. Moreover, performances that result from applying the bigram scheme are always better than with the unigram approach.

Table 3. MAP for Korean corpus (T queries)

	Mean average precision (MAP)		
Model	word	bigram (base)	HAM
PB2-nnn	0.2378	0.3729	**0.3659**
LM	0.2120*	0.3310*	0.3135*
Okapi-npn	0.2245*	0.3630*	0.3549
Lnu-ltc	0.2296	**0.3973***	0.3560
dtu-dtn	**0.2411**	0.3673*	0.3339*
atn-ntc	0.2242*	0.3270*	0.2983*
ltn-ntc	0.2370	0.3708	0.3383*
ltc-ltc	0.1606*	0.2260*	0.2299*
ntc-ntc	0.1548*	0.2506*	0.2324*
Improvement	-36.5%	0%	-6.6%

Our evaluations on the Korean collection are reported in Table 3. In this case, the best performing search model varies according to the indexing strategy. The performance differences between the best performing models ("dtu-dtn", "Lnu-ltc", PB2) are usually not significant. Using the bigram scheme as baseline, the performance differences with the word-based indexing approach were always detected as significant and in favor of the bigram approach. Comparing bigrams with the automatic decompounding strategy (under the label "HAM" in Table 3), the bigram indexing strategy tends to present a better performance, but the differences are usually not significant.

General measurements such as MAP always hide irregularities found among queries. It is interesting to note for example that for some queries, retrieval performance was poor for all search models. For example, for Topic #4 entitled "the US Secretary of Defense, William Sebastian Cohen, Beijing", the first relevant item appears in rank 37 with the PB2 model (English corpus). When inspecting top-ranked articles for this query, we found that these articles more or less contained all words included in the topic description. Moreover, their length was relatively short and these two aspects were taken into account when ranking these documents high in the response list. From a semantic point of view, these short and non-pertinent articles do not specify the reason or purpose of the visit made by the US Secretary of Defense, with content being limited to facts such

as "the US Secretary of Defense will arrive next week" or "William Sebastian Cohen will leave China tomorrow".

Topic #45 "population issue, hunger" was another difficult query. After stemming, the query is composed by the stem "hung" present in 3,036 documents, the indexing term "populat" (that occurs in 7,995 articles), and "issu" (appearing in 44,209 documents). Given this document frequency information, it would seem natural to assign more importance to the stem "hung", compared to the two other indexing terms. The term "hunger" however does not appear in any relevant document, resulting in poor retrieval performance for this query. The inclusion of the term "food" (appearing in the descriptive part of the topic) resulted in some pertinent articles being found by the search system.

6 Statistical Variations

In the previous section, we based our statistical validation on the bootstrap approach [16] in order to determine whether or not the difference between two given retrieval schemes was really significant. The null hypothesis (denoted H_0) stated that both IR systems produce the same performance level and the observed difference was simply due to random variations. To verify this assumption statistically, other statistical tests could be considered.

The first might be the Sign test [17, , pp. 157–164], in which only the direction of the difference (denoted by a "+" or "-" sign) is taken into account. This non-parametric test does not take the amount of difference into account, but only the fact that a given system performs better than the other for any given query. For example, for a set of 50 queries, System A produced better MAP for 32 queries (or 32 "+"), System B was better for 16 (or 16 "-"), and for the two remaining requests both systems showed the same performance. If the null hypothesis were true, we would expect to obtain roughly the same number of "+" or "-" signs. In the current case involving 48 experiments (the two ties results are ignored), we had 32 "+" and only 16 "-" signs. Assuming that the null hypothesis is true, the probability of observing a "+" is equal to the probability of observing a "-" (= 0.5). Thus for 48 trials the probability of observing 16 or fewer occurrences of the same sign ("+" or "-", for a two-tailed test) is only 0.0293. This value is rather small (but not null) and, in this case, when the limit was fixed at $\alpha = 5\%$, we must reject the H_0 and accept the alternative hypothesis that there were truly retrieval performance differences between System A and B.

Instead of observing only the direction of the difference between two systems, we might also consider the magnitude of the difference, not directly but by sorting them from the smallest to the largest difference. Then we could apply the Wilcoxon signed ranking test [17, pp. 352-360]. Finally, we might apply the paired t-test, a parametric test assuming that the difference between two systems follows a normal distribution. Even if the distribution of the observations was not normally shaped but the empirical distribution found to be roughly symmetric, the t-test would still be useful, given that it is a relatively robust test, in the sense that the significance level indicated is not far from the true

level. However, previous studies have shown that IR data do not always follow a normal distribution [16].

Based on 264 comparative evaluations (most of them are shown in Section 5), we applied the four statistical tests to the resultant differences. Among them for all four tests, 143 comparisons were found to be significant and 88 non-significant. Thus, for 231 (143+88) comparisons out of 264 (or 87.5%), the four tests resulted in the same decision. These four statistical tests thus are clearly in agreement, even though they use different kinds of information (e.g., for the Sign test, only the difference direction).

For the other 33 (264-231) comparisons, there was some disagreement and these cases can be subdivided into three categories. First, in 11 cases, three tests were detected to have a significant difference while the other one did not. Following inspection, we found that in 10 (out of 11) observations only the Sign test did not detect a significant difference by obtaining a p-value greater than 0.05 (see Example A in the second row of Table 4). Second, for 16 cases, two tests indicated a significant difference while the other two did not. After inspecting this sample, we found 8 observations for which both the t-test and the bootstrap detected a significant difference (see for example Case C in Table 4). In 7 other cases, both the Sign and Wilcoxon tests detected significant retrieval performance differences (see Case D in Table 4). Finally, in 6 only one test detected a significant difference while for the three others the performance difference could be due to random variations (see, for example, Case E in Table 4).

Table 4. Description and p-value for some comparisons

	Comparison	MAP	Sign test	Wilcoxon	Bootstrap	t-test
A.	ZH unigram	0.2965	**0.0595**	0.0122	0.0085	0.0084
	LM vs. ltn-ntc	0.2348	(31+ vs. 17-)			
B.	JA bigr. vs unigr.	0.1799	0.0186	0.0073	0.0430	**0.0528**
	atn-ntc vs. atn-ntc	0.1372	(32+ vs. 15-)			
C.	ZH MTools	0.3246	0.3916	0.0574	**0.0260**	**0.0299**
	PB2 vs. dtu-dtn	0.2894	(28+ vs. 21-)			
D.	JA uni+bigram	0.3026	**0.0011**	**0.0040**	0.1555	0.1740
	PB2 vs. Okapi	0.2802	(35+ vs. 12-)			
E.	KR HAM	0.3659	0.3916	**0.0297**	0.1215	0.1354
	PB2 vs. Okapi	0.3549	(28+ vs. 21-)			

To provide a more general overview of the relationship between two tests, in Figure 1 we plotted the p-values for performance comparisons from the two tests. We also computed the Pearson correlation coefficient and drew a line representing the corresponding slope. The first plot in the top left corner of Figure 1 indicates a strong correlation (r=0.9996) between the bootstrap p-values and those obtained by the t-test. Clearly, the bootstrap test agrees with the t-test results, without having to assume a Gaussian distribution.

We also tested to find out whether or not the differences distribution follows a normal distribution. In 228 (out of 264) observations, the underlying distribution

of performance difference did not follow a Gaussian distribution (Shapiro-Wilk test, significance level $\alpha = 5\%$ [18]). In both cases, the Pearson correlation coefficient between the bootstrap and t-test p-values is very high.

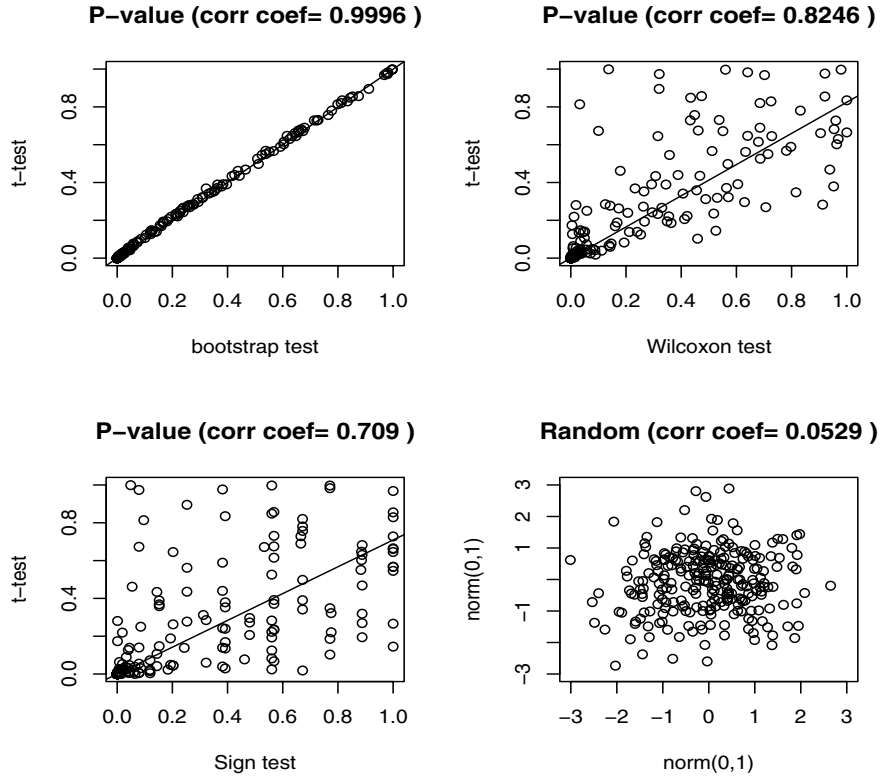

Fig. 1. Three plots of two related tests (p-values) and a random example

The relationship between the t-test and the Wilcoxon test is not as strong (top right) but still relatively high (Pearson coefficient correlation of 0.8246). When comparing p-values obtained from the t-test and the Sign test, the correlation coefficient is lower (0.709) but statistically different from 0. Finally, we plotted the same number of points obtained by generating values randomly according to the normal distribution. In this case, the true correlation coefficient is a null value, even though the depicted value is not (0.0529). The latter picture is an example of no correlation between two variables.

7 Conclusion

The experiments conducted with the NTCIR-5 test-collections show that the PB2 probabilistic model derived within the *Divergence from Randomness* framework usually produces the best mean average precision, according to different

indexing strategies and languages. For the Chinese language (Table 1), the best indexing strategy seems to be a combined approach (unigram & bigram) but when compared with a word-based approach (obtained with an automatic segmentation system), the difference is not always statistically significant.

For the Korean language, the simple bigram indexing strategy seems to be the best. When compared with the automatic decompounding strategy (HAM in Table 3), the performance difference is usually not-significant. For the Japanese language (Table 2), we may discard the unigram indexing approach, but we were not able to develop solid arguments in favor of a combined indexing approach (unigram + bigram), compared to a word-based or a simple bigram indexing scheme.

Upon analyzing the decisions that resulted from our application of a non-parametric bootstrap test, the evidence obtained strongly correlated with the (parametric) t-test conclusions. Moreover, the conclusions drawn following an application of the Wilcoxon signed ranking test correlate positively with those of the t-test. From our data, it seems that the Sign test might provide different results than the three other tests, but this divergence is not really important.

Acknowledgments. This research was supported in part by the Swiss NSF under Grant #200020-103420.

References

1. Kishida, K., Chen, K.-H., Lee, S., Kuriyama, K., Kando, N., Chen, H.-H., Myaeng, S.H.: Overview of CLIR Task at the Fifth NTCIR Workshop. In Proceedings of NTCIR-5. NII, Tokyo (2005) 1–38
2. Buckley, C., Singhal, A., Mitra, M., Salton, G.: New Retrieval Approaches using SMART. In Proceedings TREC-4. NIST, Gaithersburg (1996) 25–48
3. Singhal, A., Choi, J., Hindle, D., Lewis, D.D., Pereira, F.: AT&T at TREC-7. In Proceedings TREC-7. NIST, Gaithersburg (1999) 239–251
4. Robertson, S.E., Walker, S., Beaulieu, M.: Experimentation as a Way of Life: Okapi at TREC. Information Processing & Management **36**, (2000) 95–108
5. Amati, G., van Rijsbergen, C.J.: Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. ACM Transactions on Information Systems **20** (2002) 357–389
6. Hiemstra, D.: Using Language Models for Information Retrieval. CTIT Ph.D. Thesis (2000)
7. Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (Eds.): Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign. Lecture Notes in Computer Science, Vol. 3491. Springer, Berlin (2005)
8. Kwok, K.L. Employing Multiple Representations for Chinese Information Retrieval. Journal of the American Society for Information Science **50** (1999) 709–723
9. Luk, R.W.P., Kwok, K.L.: A Comparison of Chinese Document Indexing Strategies and Retrieval Models, ACM Transactions on Asian Languages Information Processing **1** (2002), 225–268
10. Lee, J.J., Cho, H.Y., Park, H.R.: N-gram-based Indexing for Korean Text Retrieval. Information Processing & Management **35** (1999) 427–441

11. Sproat, R.: Morphology and Computation. The MIT Press, Cambridge (1992)
12. Fujii, H., Croft, W.B.: A Comparison of Indexing Techniques for Japanese Text Retrieval. In Proceedings ACM-SIGIR. The ACM Press, New York (1993) 237–246
13. Nie, J.Y., Ren, F. Chinese Information Retrieval: using Characters or Words? Information Processing & Management **35** (1999) 443–462
14. Foo, S., Li, H.: Chinese Word Segmentation and its Effect on Information Retrieval. Information Processing & Management **40** (2004) 161–190
15. Murata, M., Ma, Q., Isahara, H.: Applying Multiple Characteristics and Techniques to Obtain High Levels of Performance in Information Retrieval. In Proceedings of NTCIR-3. NII, Tokyo (2003)
16. Savoy, J.: Statistical Inference in Retrieval Effectiveness Evaluation. Information Processing & Management **33** (1997) 495–512
17. Conover, W.J.: Practical Nonparametric Statistics. 3rd edn. John Wiley & Sons, New York (1999)
18. Maindonald, J., Braun, J.: Data Analysis and Graphics Using R. Cambridge University Press, Cambridge (2003)

Appendix: Term Weighting Formulae

In Table 5, n indicates the number of documents in the collection, t the number of indexing terms, df_j the number of documents in which the term t_j appears, the document length of D_i (the number of indexing terms) is denoted by nt_i. We assigned the value of 0.55 to the constant b, 0.1 to *slope*, while we fixed the constant k_1 at 1.2 for the English, Korean and Japanese collection and 1.0 for the Chinese corpus. For the PB2 model, we assigned $c = 3$ for the English and Korean corpus, $c = 6$ for the Japanese and $c = 1$ for the Chinese collection. These values were chosen because they usually result in improved levels of retrieval performance. Finally, the value *mean dl*, *slope* or *avdl* were fixed according to the corresponding statistics (e.g., for bigram-based indexing, 321 for ZH, 133 for JA, and 233 for KR).

Table 5. Various Weighting Schemes

ntc	$w_{ij} = \frac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t}(tf_{ik} \cdot idf_k)^2}}$	atn	$w_{ij} = idf_j \cdot \left[\frac{0.5 + 0.5 \cdot tf_{ij}}{max\ tf_{i.}}\right]$
dtn	$w_{ij} = [ln(ln(tf_{ij}) + 1) + 1] \cdot idf_j$	ltn	$w_{ij} = [ln(tf_{ij}) + 1] \cdot idf_j$
ltc	$w_{ij} = \frac{\left[ln(tf_{ij})+1\right] \cdot idf_j}{\sqrt{\sum_{k=1}^{t}([ln(tf_{ik})+1] \cdot idf_k)^2}}$	npn	$w_{ij} = tf_{ij} \cdot ln\left(\frac{n-df_j}{df_j}\right)$
dtu	$w_{ij} = \frac{\left[ln(ln(tf_{ij})+1)+1\right] \cdot idf_j}{(1-slope) \cdot pivot + (slope \cdot nt_i)}$	lnc	$w_{ij} = \frac{ln(tf_{ij})+1}{\sqrt{\sum_{k=1}^{t}(ln(tf_{ik})+1)^2}}$
Lnu	$w_{ij} = \frac{\frac{ln(tf_{ij})+1}{ln\left(\frac{l_i}{nt_i}\right)+1}}{(1-slope) \cdot pivot + (slope \cdot nt_i)}$	nnn	$w_{ij} = tf_{ij}$
Okapi	$w_{ij} = \frac{(k_1+1) \cdot tf_{ij}}{K + tf_{ij}}$ with $K = k_1 \cdot \left[(1-b) + b \cdot \frac{l_i}{avdl}\right]$		

Searching in MEDLINE: Query Expansion and Manual Indexing Evaluation

Samir Abdou, Jacques Savoy

Computer Science Department
University of Neuchatel
2009 Neuchâtel, Switzerland
{Samir.Abdou, Jacques.Savoy}@unine.ch

Abstract. Based on a relatively large subset representing one third of the MEDLINE collection, this paper evaluates ten different IR models, including recent developments in both probabilistic and language models. We show that the best performing IR models is a probabilistic model developed within the *Divergence from Randomness* framework (Amati & van Rijsbergen, 2002), which result in 170% enhancements in mean average precision when compared to the classical *tf idf* vector-space model. This paper also reports on our impact evaluations on the retrieval effectiveness of manually assigned descriptors (MeSH or Medical Subject Headings), showing that by including these terms retrieval performance can improve from 2.4% to 13.5%, depending on the underling IR model. Finally, we design a new general blind-query expansion approach showing improved retrieval performances compared to those obtained using the Rocchio approach.

Keywords. Manual Indexing, Blind Query Expansion, Medline, MeSH, Genomics TREC, Probabilistic Model, Language Model, Rocchio Query Expansion, Evaluation.

1 Introduction

MEDLINE is a well-known premier bibliographic collection that contains references to articles contained in journals on life sciences. The Genomics TREC 2004 evaluation campaign provides access to one third of this large corpus together with fifty real information need descriptions. Within this realistic context, our first goal is to evaluate the retrieval performance of various IR models, including recent developments in probabilistic and language models, and also vector-space schemes.

Second, we accept the fact that manually assigned descriptors should increase the probability of retrieving more pertinent documents, as compared to those searches based only on scientific article titles and abstracts. Manual indexing, usually based on controlled vocabularies, should prescribe a uniform and invariable choice of indexing descriptors and thus normalize orthographic variations (e.g., “database” or “data

base"), lexical variants (e.g., "analyzing," "analysis") or any other expressions having similar meanings (e.g., "computer science," "informatics").

A third issue concerns information submitted in queries to express user needs. As is commonly recognized, users do not supply all details and thus there is a lack of certain synonyms or related terms. To partially resolve this problem, a query expansion technique should take different term-term relationships into account and expand the original query. As seen from various empirical studies, this usually results in better retrieval performance.

The rest of this paper is organized as follows. Section 2 describes related works in the two different sub-domains presented in this paper: manual and automatic indexing, and automatic query expansion approaches. Section 3 depicts the main characteristics of our test-collection, while Section 4 briefly describes the IR models applied during our experiments. Section 5 explains our new query expansion model, and Section 6 evaluates the performance of various IR models, in addition to two query expansion approaches. The main findings of this paper are presented in Section 7.

2 Related Work

2.1 Manual & Automatic Indexing

Only a few studies have undertaken to directly compare the performance of manual vs. automatic indexing methods. The well-known Cranfield experiments for example studied and evaluated the retrieval impact of various manual-indexing strategies. For example, Cleverdon (1967) reported that single-word indexing was more effective than extracted terms from a controlled vocabulary, where both indexing schemes were compiled by human beings (1,400 documents, 221 queries).

In order to evaluate the importance of manually assigned descriptors, Hersh *et al.* (1994) investigated search performance differences resulting from input provided by users from different backgrounds (physicians or librarians, novices or expert users) when searching OHSUMED (a subset of the MEDLINE collection). Overall, performance differences were small and statistically insignificant, thus illustrating that the MeSH descriptors were not really advantageous. In an opposing viewpoint, Srinivasan (1996) reported that MeSH may in some cases help retrieving information in MEDLINE.

Based on the Amaryllis database, containing a French bibliographic collection (148,688 records and 25 queries), Savoy (2005) demonstrated that the inclusion of manually assigned descriptors could significantly enhance mean average precision by about 35% based on title-only queries, compared to an approach that ignored these additional descriptors. The question then arises: "Does the inclusion of MeSH headings improve mean average precision within the MEDLINE corpus?" Then, if the answer is positive: "What percentage improvement could we expect when such manually assigned descriptors are taken into account?"

2.2 Query Expansion

To provide a better match between user information needs and documents, various query expansion techniques have been suggested. The general principle is to expand the query using words or phrases having meanings similar to or related to those appearing in the original request. To achieve this, query expansion approaches consider various relationships between these words, as well as term selection mechanisms and term weighting schemes. The specific answers to these three questions may vary, leading to a variety of query expansion approaches (Efthimiadis, 1996).

In a first attempt to find related search terms, we might ask the user to select additional terms to be included in an expanded query (e.g., (Vélez *et al*., 1997)). This could be handled interactively through displaying a ranked list of retrieved items returned by the first query. Using the WordNet thesaurus, Voorhees (1994) demonstrated that terms having a lexical-semantic relation with original query words (extracted because of synonym relationship) provided very little improvement (around 1% compared to the original query).

As a second strategy for expanding the original query, Rocchio (1971) proposed taking the relevance or irrelevance of top-ranked documents into account, as indicated manually by the user. In this case, a new query would then be built automatically in the form of a linear combination of the term included in the previous query and terms automatically extracted from both relevant (with a positive weight) and non-relevant documents (with a negative weight). Empirical studies (e.g., (Salton & Buckley, 1990)) have demonstrated that such an approach is usually quite effective. Moreover, Buckley *et al*. (1996) suggested that even without looking at them or asking the user, it could be assumed that the top *k* ranked documents would be relevant. This method, denoted the pseudo-relevance feedback or blind-query expansion approach, is usually effective (at least when handling relatively large text collections).

As a third source, we might use large text corpora to derive various term-term relationships and apply statistically or information-based measures. For example, Qiu & Frei (1993) suggested that terms extracted from a similarity thesaurus that had been automatically built through calculating co-occurrence frequencies in the search collection could be added to a new query. The underlying effect was to add idiosyncratic terms to those found in underlying document collections, and related to query terms in accordance to the language being used. Kwok *et al*. (2004) suggested building an improved request by using the web to find terms related to search keywords.

In these various query expansion approaches, different underlying parameters must be specified and generally there is no single theory capable of finding the most appropriate values. Moreover, previous empirical studies conducted with newspaper corpora and more specific collections such as MEDLINE could possibly reveal other retrieval impacts that might result from the application of blind-query expansion methods. Answers to these questions can be found in the remaining sections.

3 Test Collection

The corpus used in our experiments was extracted from the MEDLINE[1] bibliographic database. It covers around 10 years of scientific publication (4,591,008 records, and around 10.6 GB of compressed data) and represents one third of the entire MEDLINE collection (approximately 13 million references).

As shown in Figure 1, each record is structured according to a specific set of fields, such as PMID (PubMed unique identifier), DP (publication date), AU (author), PT (publication type), SO (source), etc. From an IR perspective the most important sources of information include the article's title (TI), abstract (AB) and set of MeSH headings (MH) extracted from the MeSH[2] Thesaurus.

```
PMID- 10605448
…
DP  - 1978 Feb
TI  - Interrelationships between microtubules, a striated fiber, and the gametic mating structure of Chlamydomonas reinhardi
AB  - The microtubule system associated with the Chlamydomonas reinhardi flagellar apparatus is shown to differ from previous descriptions; two of the four flagellar "roots" possess only two microtubules and are associated with a finely striated fiber. In gametic cells this fiber underlies the gametic mating structure and makes contact with it. Functional interpretations are offered.
AU  - Goodenough UW
AU  - Weiss RL
PT  - Journal Article
SB  - IM
MH  - Animals
MH  - Chlamydomonas reinhardtii/*physiology/ultrastructure
MH  - Flagella/*physiology/*ultrastructure
MH  - Germ Cells/*physiology/ultrastructure
MH  - Microscopy, Electron
MH  - Microtubules/*physiology/*ultrastructure
SO  - J Cell Biol 1978 Feb;76(2):430-8.
…
```

Figure 1. Example of a MEDLINE record

This test-collection derived from the TREC 2004 evaluation campaign contains fifty topics (see examples listed in Figure 2) corresponding to real information needs expressed by biologists. Each topic is subdivided into four different fields, namely a unique identifier (<ID>), a brief title (<TITLE> or T), a full statement of the user's information need (<NEED> or N), and some background information to help assess the topic (<CONTEXT> or C). We used the three logical sections TNC to build the queries, in which the number of search terms is 13.6 in average (median: 12, standard deviation: 6.05).

[1] See http://www.nlm.nih.gov/pubs/factsheets/medline.html

[2] See the site http://www.nlm.nih.gov/mesh/meshhome.html

<ID> 2
<TITLE> Generating transgenic mice
<NEED> Find protocols for generating transgenic mice.
<CONTEXT> Determine protocols to generate transgenic mice having a single copy of the gene of interest at a specific location

<ID> 10
<TITLE> NEIL1
<NEED> Find articles about the role of NEIL in repair of DNA
<CONTEXT> Interested in role that NEIL1 plays in DNA repair.

Figure 2. Examples of topic descriptions

The relevance judgments made by two human assessors could be rated as "definitively relevant," "possibly relevant" or "not relevant." According to the Genomics TREC evaluation campaigns (Hersh *et al.*, 2005), we considered both "definitively relevant" and "possibly relevant" to be relevant items. From an inspection of the relevance assessments, the average number of relevant records per query was 165.36 (median: 115.5; standard deviation: 166.8). Query #18 or #19 had only one pertinent document while Query #42 had the greatest number of relevant articles (679).

4 IR Models

In order to obtain a broader view of the relative merit of the various retrieval models, we used seven different vector-space schemes and three probabilistic models. First, we adopted the classical *tf idf* model, wherein the weight attached to each indexing term was the product of its term occurrence frequency (or tf_{ij} for indexing term t_j in document D_i and its inverse document frequency (or idf_j). To measure similarities between documents and requests, we computed the inner product after normalizing indexing weights (model denoted "document=ntc, query=ntc" or "ntc-ntc").

Other variants might also be created, especially in cases when the occurrence of a particular term in a document is deemed a rare event. Thus, it might be good practice to assign more importance to the first occurrence of this word, compared to any successive, repeating occurrences. Therefore, the *tf* component might be computed as the ln(*tf*)+1 or as $0.5 + 0.5 \cdot [tf / \max tf \text{ in a document}]$. Of course, other weighting formulae could also be used for documents and requests, leading to different weighting combinations (see the Appendix). We might also consider that a term's presence in a shorter document would provide stronger evidence than in a longer document, leading to more complex IR models; for example the IR model denoted by "doc=Lnu" (Buckley *et al.*, 1996), "doc=dtu" (Singhal *et al.*, 1999).

In addition to these vector-space schemes, we also considered probabilistic models such as that of Okapi (Robertson *et al.*, 2000). As a second probabilistic approach, we implemented the I(*n*)B2 approach taken from the *Divergence from Randomness* (DFR) framework (Amati & van Rijsbergen, 2002), based on combining the two information measures formulated below:

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2[Prob^1_{ij}] \cdot (1 - Prob^2_{ij}) \quad (1)$$

in which $Prob^1_{ij}$ is the pure chance probability of finding tf_{ij} occurrences of the term t_j in the document D_i. On the other hand, $Prob^2_{ij}$ is the probability of encountering a new occurrence of term t_j in the document, given that we have already found tf_{ij} occurrences of this term.

Within this framework, the model I(*n*)B2 is based on the following formulae:

$$Prob^1_{ij} = \left(\frac{df_j + 0.5}{n + 1} \right)^{tfn_{ij}}$$

$$\text{and } Prob^2_{ij} = 1 - \left(\frac{tc_j + 1}{df_j \cdot (tfn_{ij} + 1)} \right) \qquad (2)$$

$$\text{with } tfn_{ij} = tf_{ij} \cdot \log_2 \left(1 + \frac{c \cdot mean\ \ dl}{l_i} \right)$$

where df_j indicates the number of documents indexed with the term t_j, tc_j the number of occurrences of term t_j in the collection, *n* the number of documents in the corpus, *mean dl* (= 146) the average document length, and *c* a constant (fixed at 1.5).

Finally, we also considered an approach based on a language model (LM) (Hiemstra, 2000), known as a non-parametric probabilistic model (while the Okapi and I(*n*)B2 are parametric models). Thus the estimations needed would not be based on any known term distribution but rather directly from its occurrence frequencies in the document D or the corpus C. From the language model paradigm, we might consider various implementations and smoothing methods. In this study, we adopted a model proposed by Hiemstra (2000; 2002), as described in Equation 3, which combines an estimation based on the document ($Prob[t_j \mid D_i]$) and on the corpus ($Prob[t_j \mid C]$)

$$Prob[D_i \mid Q] = Prob[D_i] \cdot \prod_{tj \in Q} [\lambda_j \cdot Prob[t_j \mid D_i] + (1-\lambda_j) \cdot Prob[t_j \mid C]]$$

$$\text{with } Prob[t_j \mid D_i] = \left(\frac{tf_{ij}}{l_i} \right)$$

$$\text{and } Prob[t_j \mid C] = \left(\frac{df_j}{lc} \right) \quad \text{with } lc = \sum_{k=1}^{t} df_k \qquad (3)$$

in which λ_j is a smoothing factor (fixed at 0.35 for all term t_j) and *lc* the size of the corpus C, measured by the number of occurrences of each of the *t* terms included in the inverted file.

5 Blind-Query Expansion

Various general query expansion approaches have been suggested and in this paper we compared ours with that of Rocchio (1971; Buckley *et al*., 1996). In this latter case, the system was allowed to add *m* terms extracted from the *k* best-ranked documents from the original query. Each new query was derived by applying the following formula:

$$w'_j = \alpha \cdot w_j + (\beta/k) \cdot \sum_{j=1}^{k} w_{ij} \quad (4)$$

in which w'_j denotes the weight attached to the *j*th query term, based on the weight of this term in the previous query (denoted by w_j), and w_{ij} the indexing term weight attached to this *j*th term in the document D_i appearing in the top *k* ranks. In our evaluation, we fixed $\alpha = 2.0$, $\beta = 0.75$.

To define our new blind-query expansion denoted "IDF Query Expansion" (or simply IDFQE), we adopted the following procedure. First we formed the root set of search terms composed of all terms included in the original query Q and all indexing terms appearing in the *k* best ranked documents. The weight attached to each term in this root set was computed as follows:

$$w'_j = \alpha \cdot I_Q(tj) \cdot tf_j + \left(\frac{\beta}{k}\right) \cdot \sum_{i=1}^{k} I_{D_i}(t_j) \cdot idf_j \quad (5)$$

$$\text{with } I_Q(t_j) = 1 \text{ if } t_j \in Q,\ 0 \text{ otherwise}, \quad I_{D_i}(t_j) = 1 \text{ if } t_j \in D_i,\ 0 \text{ otherwise}$$

where for term t_j, $idf_j = \log(n/df_j)$ (or the classical idf value) and $I_Q(t_j)$ (or $I_{Di}(t_j)$) is an indicator function returning the value 1 if the term t_j belongs to the query Q (or the document D_i), otherwise 0. In this weighting scheme, if a term appears only in the original query Q, its weight would be $\alpha \cdot tf_j$, while a term appearing in only one document would have a weight of $(\beta/k) \cdot idf_j$.

The root set elements were then sorted in descending order according to their weight. To form the new query Q', we selected the top *m* search terms, and the weights attached to these selected terms in the new query being the same as those used in the root set. We thus used the same weighting scheme to select and weight the new search terms.

6 Evaluation

To evaluate our various IR schemes, we adopted the mean average precision (MAP) to measure retrieval performance (based on 1,000 records). To statistically determine whether or not a given search strategy would be better than another, we applied the bootstrap methodology (Savoy, 1997). In the statistical tests we applied, the null hypothesis H_0 states that both retrieval schemes produce similar MAP performance. Such a null hypothesis would be accepted if two retrieval schemes returned statistically a similar MAP, otherwise it would be rejected. Thus, in the tables included in this paper, we have underlined any statistically significant differences resulting from a two-sided non-parametric bootstrap test (significance level α=5%).

6.1 IR Models Evaluation

Table 1 depicts the MAP for the MEDLINE collection subset obtained using the Porter (1980) stemmer with different IR models. In this evaluation we used the documents' title and abstract together with their MeSH descriptors. This table depicts the best

performances under a given condition bold, and these values are then used as the baseline for statistical testing. The first column indicates the tested IR models and in the second column the evaluation of the corresponding models.

Model	MAP (% change)
I(n)B2-nnn	**0.3810** (+174%)
LM	0.3630 (+161%)
Okapi-npn	0.3573 (+156%)
Lnu-ltc	0.2962 (+113%)
dtu-dtn	0.3402 (+144%)
atn-ntc	0.3192 (+129%)
ltn-ntc	0.3098 (+122%)
lnc-ltc	0.1906 (+37%)
ltc-ltc	0.1948 (+40%)
ntc-ntc (*tf idf*)	0.1393

Table 1. Mean average precision (MAP) of different IR models and percentage of change over the classical *tf idf* model

The evaluations shown in Table 1 indicate that the I(*n*)B2 probabilistic model produces the best retrieval performance (baseline). The differences between the I(*n*)B2 on the one hand, and the language model (LM) or Okapi models on the other are not statistically significant. With the other seven vector-space IR models the MAP differences are always statistically significant, showing an improvement of around 170% over the classical *tf idf* IR model (denoted "ntc-ntc" in Table 1).

6.2 Manually Assigned Headings

In Table 2, we listed the mean average precision achieved when the best performing IR models used only the article's title and abstract to build the document surrogate. Overall, retrieval performance under this indexing restriction was lower than the corresponding system using manually assigned descriptors (shown under the label "with MeSH"). Average increases were around 8.5% when the search system included the MeSH headings. The greatest enhancement however was obtained with the Okapi model (from 0.3217 to 0.3573, an absolute improvement of 0.035, or a relative increase of 11.1%). When comparing the MAP before and after including MeSH descriptors, performance differences were statistically significant for the best performing IR systems (the first four IR models depicted in Table 2).

	Mean average precision (MAP)	
Model	With MeSH	Without MeSH
I(*n*)B2-nnn	**0.3810** (+8.4%)	0.3516
LM	0.3630 (+9.6%)	0.3311
Okapi-npn	0.3573 (+11.1%)	0.3217
Lnu-ltc	0.2962 (+10.0%)	0.2693
dtu-dtn	0.3402 (+4.8%)	0.3245
atn-ntc	0.3192 (+2.4%)	0.3117
ntc-ntc	0.1393 (+13.5%)	0.1227

Table 2. MAP with and without MeSH headings

6.3 Evaluating Query Modification and Expansion

To evaluate both our new blind query expansion method and the Rocchio scheme, we used the best performing model, namely I(*n*)B2 as shown in Table 3. In the bottom part of this table, the rows starting with "*k* doc/*m* terms" indicate the number of top-ranked documents and the number of terms used to enlarge the original query. We then compared the MAP achieved by the two query expansion approaches using either the Rocchio or IDFQE strategy

	Mean average precision (MAP)	
I(*n*)B2	0.3810	
Query expansion	IDFQE	Rocchio
3 docs / 10 terms	0.3640	0.2840
3 docs / 20 terms	0.3599	0.2819
5 docs / 10 terms	0.3860	0.3011
5 docs / 20 terms	0.3833	0.3039
10 docs / 10 terms	**0.3976**	0.3282
10 docs / 20 terms	0.3896	0.3257
10 docs / 30 terms	0.3881	0.3243

Table 3. MAP achieved by various query expansion models

The data in Table 3 indicates that the Rocchio query expansion approach provided lower MAP than did the baseline performance (0.3810, MAP before query expansion), and that these differences are always statistically significant. Peat & Willett (1991) provide one explanation for the poor performance shown by the Rocchio approach. In their study they show that query terms have a greater occurrence frequency than other terms. Second, query expansion approaches based on term co-occurrence data will include additional terms that also have a greater occurrence frequency in the documents. In such cases, these additional search terms will not prove effective in discriminating between relevant and non-relevant documents. In such circumstances, the final effect on retrieval performance could be negative.

On the other hand, our suggested *idf*-based query expansion (IDFQE) tends to produce better MAP than the baseline and offers clearly better retrieval performance than

the Rocchio scheme. When comparing our IDFQE scheme with the baseline (0.3810), the performance differences are usually not significant. However, when comparing Rocchio with our IDFQE query expansion model, the performance differences were statistically significant, and always in favor of the IDFQE.

7 Conclusion

Experiments conducted on a large subset of the MEDLINE collection show that the best mean average precision is obtained using the I(*n*)B2 probabilistic model (see Table 1). Moreover, when compared to various vector-space models, the performance differences are always statistically significant and in favor of the I(*n*)B2 model. When compared to the Okapi or a language model however, the performance differences are not statistically significant.

Including the Medical Subject Headings (MeSH) when indexing scientific articles improves retrieval performance by between 2.4% and 13.5%, depending on the underlying IR model (see Table 2). These differences are usually statistically significant.

In this paper, we also proposed a new query expansion approach using the *idf* values to weight the new search terms. Depending on the parameter setting, the resulting improvement achieved by this new query expansion varied, but the retrieval performances for this new blind query expansion are statistically better than the MAP obtained by the Rocchio scheme.

Although the focus of this study was on retrieval effectiveness for meta-information sources, the effectiveness of manually assigned descriptors could of course be studied in a variety of other perspectives. One possible research objective might be to build and evaluate the most effective means of automatically assigning MeSH (or concept-based indexing) terms to free text expressions, including both scientific articles and queries (see, for example, Kim *et al*. (2001), Mao & Chu (2002), Chu *et al*. (2003)). The use of MeSH terms or generally controlled vocabulary system in automatic indexing is however questionable in some cases due to its inadequate specificity (Cimino, 1995) or the presence of redundant terms (e.g., two entries for the same diagnosis in a given thesaurus). On the other hand, such automatic assignment could also be based on citation patterns found in scientific articles, under the assumption that these articles share some semantic content with the citing document (e.g., using a bibliographic coupling measure).

Moreover, given that MeSH is a hierarchical thesaurus, users looking for certain information might use various term relationships to build either a more specific or a broader topic, through including more specific or more general concepts. Moreover, such query expansion could be done automatically based on relationships such as meronym (*part of*) or hyponym/hypernym (*kind of*, as used in "a plant is hypernymic to flower" and "rose is hyponymic to flower"). Such a task may however be more complex than it appears. In fact, search concepts often fall into more than one class and finding related terms may be no more trivial when applying strict hierarchies (Cimino, 1995). For example, terms related to the concept of "lung cancer" may appear in the "Lung Disease," "Cancer," "Tumor," or "Pulmonary Neoplasms" class. As

is evidenced here there are several avenues for further research on the general theme of controlled vocabularies and on more general knowledge-based systems working in specific domains.

Acknowledgments
This research was supported in part by the Swiss NSF under Grants #200020-103420 and #200021-113273.

References

Amati, G., & van Rijsbergen, C.J. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM-Transactions on Information Systems*, 20(4), 2002, 357-389.

Buckley, C., Singhal, A., Mitra, M., & Salton, G. New retrieval approaches using SMART. In *Proceedings of TREC-4*. Gaithersburg (MA), 1996, 25-48.

Chu, W. W., Liu V. Z., & Mao, W. Techniques for textual document indexing and retrieval via knowledge sources and data mining. In W. Wu, H. Xiong, S. Shekhar (Eds), *Clustering & Information Retrieval*, Kluwer, 2003, 135-160.

Cimino, J. J. Vocabulary and health care information technology: State of the art. *Journal of the American Society for Information Science*, 46(10), 1995, 777-782.

Cleverdon, C.W. The Cranfield tests on index language devices. In *ASLIB Proceedings*, 19, 1967, 173-192.

Efthimiadis, E.N. Query expansion. *Annual Review of Information Science and Technology*, 31, 1996, 121-187.

Hersh, W., Buckley, C., Leone, T.J., & Hickam, D. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the ACM-SIGIR*. Dublin, 1994, 192-201.

Hersh, W.R., Bhuptiraju, R.T., Ross, L., Johnson, P., Cohen, A.H., & Kraemer, D.F. TREC 2004 genomics track overview. In *Proceedings TREC-2004*. Gaithersburg, 2005, 192-201.

Hiemstra, D. Using language models for information retrieval. CTIT Ph.D. Thesis, 2000.

Hiemstra, D. Term-specific smoothing for the language modeling approach to information retrieval. The importance of a query term. In *Proceedings of the ACM-SIGIR*, Tempere, 2002, 35-41.

Kim, W., Aronson, A. R., & Wilbur, W. J. Automatic MeSH term assignment and quality assessment. In *Proceedings AMIA*, 2001, 319-323.

Kwok K.L., Grunfield, L, Sun, H.L., & Deng, P. TREC 2004 robust track experiments using PIRCS. In *Proceedings TREC 2004*, Gaithersburg (MD), 2004.

Mao, W., & Chu, W. W. Free-text document retrieval via phrase-based vector space model. In *Proceedings AMIA*, 2002, 489-493.

Peat, H. J., & Willett, P. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5), 1991, 378-383.

Porter, M.F. An algorithm for suffix stripping. *Program*, 14(3), 1980, 130-137.

Qiu, Y., & Frei, H.P. Concept based query expansion. In *Proceedings of the ACM-SIGIR*. Pittsburgh (PA), 1993, 160-169.

Robertson, S.E., Walker, S., & Beaulieu, M. Experimentation as a way of life: Okapi at TREC. *Information Processing &Management*, 36(1), 2000, 95-108.

Rocchio, J.J.Jr.: Relevance feedback in information retrieval. In G. Salon (Ed.), *The SMART Retrieval System.* Prentice-Hall Inc., Englewood Cliffs (NJ), 1971, 313-323.
Salton, G., & Buckley, C. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4), 1990, 288-297.
Savoy, J. Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, 33(4), 1997, 495-512.
Savoy, J. Bibliographic database access using free-text and controlled vocabulary: An evaluation. *Information Processing & Management*, 41(4), 2005, 873-890.
Singhal, A., Choi, J., Hindle, D., Lewis, D.D. & Pereira, F. AT&T at TREC-7. In *Proceedings TREC-7*, Gaithersburg (MA), 1999, 239-251.
Srinivasan, P. Optimal document-indexing vocabulary for Medline. *Information Processing & Management*, 32(5), 1996, 503-514.
Vélez, B., Weiss, R., Sheldon, M.A., & Gifford, D.K. Fast and effective query refinement. In *Proceedings of the ACM-SIGIR*. Philadelphia (PA), 1997, 6-15.
Voorhees, E.M. Query expansion using lexical-semantic relations. In *Proceedings of the ACM-SIGIR*. Dublin, 1994, 61-69.

Appendix: Term weighting formulae

In Table 4, *n* indicates the number of documents in the collection, *t* the number of indexing terms, df_j the number of documents in which the term t_j appears, for D_i, document length (the number of indexing terms) is denoted by nt_i, and *avdl*, *b*, k_1, *pivot* and *slope* are constants (fixed at $b = 0.55$, $k_1 = 1.2$, *avdl* = 146).

atn	$w_{ij} = idf_j \cdot [0.5 + 0.5 \cdot tf_{ij} / \max tf_{i.}]$	nnn	$w_{ij} = tf_{ij}$
dtu	$w_{ij} = \frac{\left(\ln\left(\ln(tf_{ij}) + 1\right) + 1\right) \cdot idf_j}{(1 - slope) \cdot pivot + slope \cdot nt_i}$	dtn	$w_{ij} = [\ln(\ln(tf_{ij}) + 1) + 1] \cdot idf_j$
Okapi	$w_{ij} = \frac{\left((k_1 + 1) \cdot tf_{ij}\right)}{\left(K + tf_{ij}\right)}$	ltn	$w_{ij} = (\ln(tf_{ij}) + 1) \cdot idf_j$
Lnu	$w_{ij} = \frac{\left(1 + \ln(tf_{ij}) \,\middle/\, \ln(\text{mean } tf) + 1\right)}{(1 - slope) \cdot pivot + slope \cdot nt_i}$	npn	$w_{ij} = tf_{ij} \cdot \ln[(n - df_j) / df_j]$
lnc	$w_{ij} = \frac{\ln(tf_{ij}) + 1}{\sqrt{\sum_{k=1}^{t} (\ln(tf_{ik}) + 1)^2}}$	ntc	$w_{ij} = \frac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t} (tf_{ik} \cdot idf_k)^2}}$
ltc	$w_{ij} = \frac{\left(\ln(tf_{ij}) + 1\right) \cdot idf_j}{\sqrt{\sum_{k=1}^{t} \left(\left(\ln(tf_{ik}) + 1\right) \cdot idf_k\right)^2}}$		

Table 4. Weighting schemes

Considérations sur l'évaluation de la robustesse en recherche d'information

Samir Abdou, Jacques Savoy

Institut d'informatique
Université de Neuchâtel, rue Emile Argand 11, 2009 Neuchâtel (Suisse)
Samir.Abdou@unine.ch, Jacques.Savoy@unine.ch

RÉSUMÉ. *Cette communication évalue et compare l'efficacité de modèles vectoriels, probabilistes ou de langue afin de dépister des articles de presse rédigés en langue française. En se basant sur un corpus créé durant trois campagnes d'évaluation CLEF et comprenant 151 requêtes, nous avons pu découvrir les raisons expliquant la faible performance des divers modèles face à des requêtes difficiles. L'évaluation de la robustesse de ces approches s'avère tout de même peu aisée car la moyenne arithmétique (MAP) ou la moyenne géométrique (GMAP) ne présentent pas toutes les caractéristiques souhaitables. Afin de compléter ces deux mesures, nous proposons de recourir au score du premier document dépisté (FRS). Nous avons comparé les résultats de ces trois mesures de performance en particulier avec l'expansion aveugle des requêtes.*

ABSTRACT. *This paper describes and evaluates vector-space, probabilistic and language IR models used to retrieve news articles from a corpus written in the French language. Based on three CLEF test-collections and 151 topics, we analyze the retrieval effectiveness of these approaches and analyze the poor retrieval results of hard topics. An appropriate robust evaluation is not easy because both the mean average precision (MAP) or the geometric mean (GMAP) present some drawbacks. In order to obtain a better picture, we suggest using the First Relevant Score (or FRS, based on the rank of the first relevant item). We evaluate and compare these three measures in particular when using blind query expansion technique.*

MOTS-CLÉS : *Evaluation de recherche robuste ; expansion aveugle ; requêtes difficiles.*

KEY WORDS: *Robust evaluation; blind query expansion; hard queries.*

1. Introduction

La recherche d'information propose continuellement des améliorations de ses stratégies d'indexation et de dépistage et, avec les années, les progrès s'accumulent. Dès lors on peut penser que les modèles les plus récents arriveront à dépister au moins une bonne réponse et à la présenter parmi les dix premières références, quelle

que soit la requête soumise. Cet apriori est conforté par le fait que la grande majorité des personnes interrogées s'avèrent satisfaites d'un moteur comme *Google*.

Pour un service commercial, l'absence de bonne réponse parmi les dix premières références retournées crée un impact négatif important. En effet, on sait qu'un client insatisfait en parlera à environ 25 autres personnes tandis que, dans le cas contraire, il aura en moyenne l'occasion de discuter d'une bonne expérience avec cinq autres personnes. Pouvoir garantir un service minimum (par exemple, retourner au moins une bonne réponse parmi les dix premières références dépistées) est un critère qu'un service enligne souhaite atteindre, quitte à renoncer à disposer d'une précision moyenne élevée. Afin d'analyser empiriquement cette question, la piste « robuste » a été créée lors des campagnes d'évaluation TREC depuis 2003 [VOO 05 ; 06] et plus récemment dans le cadre de CLEF (depuis 2006).

Dans cette communication, nous désirons présenter et mesurer l'effet de quelques stratégies pouvant améliorer le dépistage de documents pertinents pour les requêtes ardues. Nous souhaitons également limiter notre champ d'investigation à une seule langue soit le français d'une part et, d'autre part, à des requêtes courtes correspondant mieux à la réalité, celle du *Web* pour le moins. Face à des requêtes plus longues, le système de dépistage peut cerner avec plus de précision le véritable centre d'intérêt de l'usager voire de lever les ambiguïtés d'un mot ou groupe de mots. Par exemple, une interrogation limitée au mot « chat » peut évoquer un animal domestique, une messagerie instantanée, un espace culturel, un titre d'ouvrage, un acronyme, etc. Pour d'autres langues, ce mot dispose d'un espace sémantique plus restreint ; ce terme fut la requête la plus fréquente sur *Yahoo.es* durant l'année 2006 (la deuxième sur *Yahoo.it* et la cinquième sur *Yahoo.de*).

En premier lieu, signalons que la détection des requêtes ardues ne peut pas s'opérer de manière intrinsèque. La simple lecture d'une interrogation ne suffit pas, même à un être humain, pour la classer, de manière fiable, dans la catégorie des requêtes faciles ou, au contraire, dans celles des difficiles [VOO 05 ; 06]. Comment différencier les requêtes « Les succès d'Ayrton Senna » (n° 121), « Le mariage Jackson-Presley » (n° 123), ou la demande « Traité de paix de Dayton » (n° 197) de la requête « Forces de maintien de la paix en Bosnie » (n° 48) ? Dans ces exemples, seule la dernière s'avère difficile pour toutes les stratégies de dépistage. Cette distinction entre interrogations difficiles et les autres doit s'appuyer sur la collection de documents. Pourtant, même avec cette information supplémentaire, les divers systèmes automatiques n'arrivent pas, de manière fiable, à prédire le degré de difficulté d'une requête [VOO 05 ; 06], [CAR 05 ; 06].

Deuxièmement, le nombre restreint de bonnes réponses ne peut pas être vu comme une indication précise de la difficulté d'une requête. On peut imaginer qu'une interrogation disposant d'une seule, voire de deux ou trois bonnes réponses perdues dans une collection volumineuse de documents serait un indice fiable de la difficulté sous-jacente de la requête. En se basant sur les exemples précédents, on peut constater que la réponse adéquate pour une requête peut être aisée même si

cette dernière possède un nombre restreint de bonnes réponses. Ainsi, la précision moyenne est maximale (1,0 avec le modèle Okapi) pour les requêtes n° 121 (« Les succès d'Ayrton Senna ») ayant une seule bonne réponse ou pour la demande n° 123 (« Le mariage Jackson-Presley »), avec deux bonnes réponses. Par contre, pour la requête n° 155 « Les risques du téléphone portable » possédant seulement deux documents pertinents, la précision moyenne s'élevait seulement à 0,0082. Pour la requête n° 197 (« Traité de paix de Dayton »), la précision moyenne était élevée (soit 0,7762) malgré les 131 documents pertinents.

Comme ces premières explications s'avèrent insuffisantes, nous avons repris un corpus d'articles de presse écrits en langue française (voir section 2) pour étudier un groupe de requêtes difficiles. Afin de travailler avec les meilleures stratégies de dépistage, nous avons décidé d'implémenter le modèle Okapi, deux approches tirées de la famille "*Divergence from Randomness*" et un modèle de langue (voir section 3). Deux modèles vectoriels compléteront cette liste afin d'obtenir un panorama plus complet. La section 4 aborde la question de l'évaluation et permet d'avoir une vue plus critique sur des mesures comme la moyenne arithmétique des précisions moyennes ou la moyenne géométrique. La section 5 analyse la stratégie de l'expansion aveugle des requêtes et démontre que les diverses mesures d'évaluation apportent des conclusions différentes.

2. Le corpus d'évaluation

Afin d'étudier les problèmes sous-jacents de l'évaluation, en particulier en face de requêtes difficiles, nous avons eu recours à la collection de documents proposée dans la piste robuste de la campagne d'évaluation CLEF-2006. Ce corpus comprend les collections en langue française utilisées lors des campagnes des années 2001 [PET 02], 2002 [PET 03] et 2003 [PET 04] et donc un ensemble relativement important de requêtes. Ce corpus comprend des articles de presse du journal *Le Monde* (1994), et des dépêches d'agence provenant de l'*Agence Télégraphique Suisse* ou *ATS* (1994-1995).

	2001	2002	2003
Source	*Le Monde* 94 ATS 94	*Le Monde* 94 ATS 94	*Le Monde* 94 ATS 94 & 95
Taille	243 MB	243 MB	331 MB
No. docs	87 191	87 191	129 806
Requête	n° 41 à n° 90	n° 91 à n° 140	n° 141à n° 200

***Table 1** : Quelques statistiques sur les quatre corpus*

Comme l'illustrent les données de la table 1, les mêmes documents sont repris dans les années 2001 et 2002. De plus, ces articles sont extraits de la même année

(1994) et couvrent des nouvelles politiques, économiques, sociales mais également des événements sportifs ou scientifiques. Lors de la campagne d'évaluation CLEF 2003, 42 615 documents provenant de l'ATS en 1995 ont été ajoutés au corpus. Au niveau des interrogations, les requêtes n° 41 à n° 140 possèdent des documents pertinents dans les sources extraites de l'année 1994 tandis que les bonnes réponses pour les 60 dernières demandes (du n° 141 à n° 200) doivent être recherchées dans les années 1994 et 1995.

<TOP> <NUM> C094 </NUM>
<TITLE> Le retour de Soljénitsyne </TITLE>
<DESC> Trouver les documents qui traitent du retour en Russie du prix Nobel de littérature, Soljénitsyne. </DESC>
<NARR> Les documents pertinents donneront les raisons et la date du retour de Soljénitsyne en Russie. Ils pourront également mentionner les raisons de son émigration aux Etats-Unis. </NARR>
</TOP>

<TOP> <NUM> C156 </NUM>
<TITLE> Les syndicats en Europe </TITLE>
<DESC> Quelles sont les différences dans le rôle et l'importance des syndicats entre les pays européens? </DESC>
<NARR> Les documents pertinents doivent comparer le rôle, le statut ou l'importance des syndicats entre deux ou plusieurs pays européens. Les informations pertinentes incluront le niveau d'organisation, les mécanismes de négociations salariales, et le climat général du marché du travail. </NARR>
</TOP>

<TOP> <NUM> C200 </NUM>
<TITLE> Innondationeurs en Hollande et en Allemagne </TITLE>
<DESC> Trouvez des statistiques sur les inondations en Hollande et en Allemagne en 1995 </DESC>
<NARR> Les documents pertinents mesureront les effets des dommages causés par l'inondation qui a eu lieu en Allemagne et en Hollande en 1995 en termes de nombres de personnes et d'animaux évacués et/ou de pertes économiques </NARR>
</TOP>

Table 2 : *Exemples de requêtes du corpus*

Suivant le modèle des campagnes TREC, chaque requête possède principalement trois champs logiques, à savoir un titre bref (<TITLE> ou T), une phrase décrivant le besoin d'information (<DESC> ou D) et une partie narrative (<NARR> ou N) spécifiant plus précisément le contexte de la demande ainsi que des critères de pertinence permettant de mieux évaluer les articles dépistés. La table 2 présente quelques exemples. Pour l'essentiel de nos évaluations, nous avons retenu uniquement la partie "titre" (T) pour construire les requêtes. Avec cette limite, la longueur moyenne des requêtes s'élève à 2,91 termes d'indexation tandis que le

recours aux deux champs “titre” et “descriptif” (TD) produisent une longueur moyenne de 7,51 mots.

Les thèmes de ces requêtes couvrent des domaines variés comme “Des pesticides dans la nourriture pour bébés”, “El Niño et le temps”, “Embargo sur l'Iraq” ou “La vache folle en Europe”. Elles incluent tant des questions ayant un intérêt plutôt régional (“Initiative suisse pour les Alpes” ou “La querelle bavaroise sur les crucifix”), un focus national (“L'affaire du sang contaminé”, “Les affaires en France”) ou une couverture internationale (“La sonde spatiale Ulysse” ou “ONU / Etats-Unis invasion d'Haïti”).

Si l’on analyse les jugements de pertinence, on remarque que neuf requêtes ne possèdent pas de bonnes réponses dans le corpus (soit n° 64, n° 146, n° 160, n° 161, n° 166, n° 169, n° 172, n° 191 et n° 194). Nos évaluations porteront donc sur 151 requêtes. Sur cet ensemble, le nombre moyen d’articles pertinents par requête s’élève à 23,45 (médiane: 13, minimum: 1, maximum: 193 (n° 181 “Essais nucléaires français”) et écart-type: 31.04).

3. Les stratégies d'indexation et modèles de dépistage

Nous désirons obtenir une vision assez large de la performance de divers modèles de dépistage de l'information afin de pouvoir fonder nos conclusions sur de solides bases. Dans ce but, nous avons indexé les documents (et les requêtes) selon la formulation classique $tf \cdot idf$, c’est-à-dire en tenant 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’$idf_j = \log(n/df_j)$). Cette pondération a été normalisée par la formule du cosinus.

D'autres variantes du modèle vectoriel ont été proposées comme, par exemple, le recours au logarithme afin d’imposer que la première occurrence d'un terme possède plus d'influence (e.g., $\log(tf)+1$) ou que la longueur du document soit prise en compte. Dans cet article, nous avons repris le modèle “Lnu“ [BUC 96] correspondant à la formule suivante :

$$w_{ij} = [(\ln(tf_{ij})+1) / (\ln(mean\ tf_{i.})+1)] \ / \ [(1\text{-}slope) \cdot pivot + slope \cdot nt_i] \qquad (1)$$

dans laquelle w_{ij} indiquant le poids du j^e terme dans la représentation du i^e document, nt_j la longueur du i^e document (le nombre de termes d’indexation distincts), *slope* une constante (fixée à 0,1 dans nos évaluations) et *pivot*, constante fixée à 118.

En plus de ces deux modèles basés sur la vision géométrique du modèle vectoriel, nous avons considéré le modèle probabiliste Okapi [ROB 00] utilisant la formulation suivante :

$$w_{ij} = [(k_1+1) \cdot tf_{ij}] / (K + tf_{ij}) \quad \text{avec } K = k_1 \cdot [(1\text{-}b) + ((b \cdot l_i) / mean\ dl)] \qquad (2)$$

dans laquelle l_i est la longueur du i^e article (mesurée en nombre de termes d'indexation), et b, k_1, *mean dl* des constantes fixées à $b = 0{,}4$, $k_1 = 1{,}2$ et *mean dl* = 180.

Comme deuxième modèle probabiliste, nous avons implémenté le modèle I(n_e)C2, un des membres de la famille *Divergence from Randomness* (DFR) [AMA 02]. Dans ce dernier cas, la pondération w_{ij} combine deux mesures d'information, à savoir :

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = Inf^1_{ij} \cdot (1 - Prob^2_{ij})$$
$$Prob^2_{ij} = 1 - [(tc_j + 1) / (df_j \cdot (tfn_{ij} + 1))] \quad \text{avec } tfn_{ij} = tf_{ij} \cdot \ln[1 + ((c \cdot mean\ dl) / l_i)]$$
$$Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (n_e + 0{,}5)] \quad \text{avec } n_e = n \cdot [1 - [(n-1)/n]^{tc_j}] \qquad (3)$$

dans laquelle tc_j représente le nombre d'occurrences du j^e terme dans la collection. Ce modèle a été conçu pour apporter une meilleure réponse face à des requêtes difficiles [PLA 05].

On remarquera que ce dernier modèle dispose d'un paramètre (noté c) que l'on doit fixer plus ou moins arbitrairement ou selon la performance obtenue sur d'anciennes requêtes. Afin d'éviter de devoir inclure de tels paramètres, Amati [AMA 06] propose un nouveau modèle nommé DLH et dérivé de la famille DFR. Comme l'indique l'équation suivante, cette approche ne dispose d'aucun paramètre sous-jacent.

$$w_{ij} = [\, tf_{ij} \cdot \log_2(p_{ij} / pc_j) + 0{,}5 \cdot \log_2[2 \cdot \pi \cdot tf_{ij} \cdot (1 - p_{ij})]\,] \;/\; [\, tf_{ij} + 1\,]$$
$$\text{avec } p_{ij} = tf_{ij} / nt_i \quad \text{et } pc_j = tc_j / (n \cdot mean\ dl) \qquad (4)$$

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 ou dans le corpus C. Dans cet article, nous avons repris le modèle de Hiemstra [HIE 00] décrit dans l'équation 5 et qui combine une estimation basée sur le document (soit Prob[t_j | D_i]) et sur le corpus (Prob[t_j | C]).

$$Prob[D_i \mid Q] = Prob[D_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot Prob[t_j \mid D_i] + (1-\lambda_j) \cdot Prob[t_j \mid C]] \qquad (5)$$
$$\text{avec } Prob[t_j \mid D_i] = tf_{ij} / nt_i \quad \text{et } Prob[t_j \mid C] = df_j / lc \quad \text{avec } lc = \sum_k df_k \qquad (6)$$

dans laquelle λ_j est un facteur de lissage (une constante pour tous les termes t_j, et qui est fixée à 0,35) et *lc* indique la taille du corpus C.

Lors de l'indexation, les mots les plus fréquents ou appartenant à une forme grammaticale peu intéressante (conjonction, préposition, pronom, déterminant) sont éliminés (soit 463 mots dans nos évaluations). De même, nous procédons à la suppression automatique des suffixes liés à la flexion (pluriel, féminin) ainsi qu'à

quelques formes liées à la dérivation morphologique (par exemple "-esse" ou "-ique" dans "volcanique")[1] [SAV 02].

4. Evaluation et ses lacunes

Afin de mesurer la performance de ces divers modèles de dépistage, nous avons utilisé la précision moyenne (PM) pour chaque requête, valeur calculée par le logiciel `trec_eval`. Cette mesure a été adoptée par diverses campagnes d'évaluation pour évaluer la qualité de la réponse à une interrogation. Elle possède l'avantage de tenir compte de la précision, du rappel et du rang des documents pertinents dépistés. Pourtant cette mesure ainsi que la moyenne arithmétique de ces précisions (MAP) soulèvent quelques interrogations lorsque l'on étudie quelques requêtes comme le montre la section 4.1. Comme alternative et désirant accorder plus d'influence aux requêtes difficiles, nous pouvons recourir à la moyenne géométrique de ces précisions (GMAP) comme le démontre la section 4.2. La section 4.3 analyse les requêtes difficiles de notre corpus. La section 4.4. propose une mesure complémentaire pour évaluer les améliorations possibles touchant en particulier les requêtes difficiles.

4.1. La précision moyenne et la MAP

Afin de connaître la performance que l'on peut associer à une requête, la communauté scientifique a adopté comme mesure principale la précision moyenne (PM). Son calcul s'opère selon le principe suivant. Pour chaque requête, on détermine la précision après chaque document pertinent, puis on calcule une moyenne arithmétique sur l'ensemble de ces valeurs. Si une interrogation ne dépiste aucun document pertinent, sa précision moyenne sera nulle. Dans la table 3, la précision moyenne de la requête A possédant trois documents pertinents s'élève à $(1/3)\cdot(1/2 + 2/3 + 3/35) = 0{,}4175$.

Pourtant la précision moyenne (PM) possède quelques inconvénients. En premier lieu cette valeur reste difficile à interpréter pour un usager. Que signifie une précision moyenne de 0,3 ? Ce n'est pas la précision après 5 ou 10 documents dépistés, valeur qui serait simple à interpréter pour l'utilisateur. Deuxièmement, comme l'illustre la table 3, des différences de précision moyenne importantes comme par exemple 0,6759 vs. 0,4175 (variation relative de 60 %) ne semblent pas correspondre à une différence aussi significative pour un usager. En effet, le classement proposé par la requête A ne s'éloigne pas beaucoup de la liste obtenue avec la requête B. En tout cas, l'usager n'attribuerait pas à cette variation une amplitude aussi élevée que 60 %.

[1] La liste de mots-outils et l'enracineur sont disponible sur le site www.unine.ch/info/clef/

Rang	Requête A	Requête B
1	NP	P 1/1
2	P 1/2	P 2/2
3	P 2/3	NP
…	NP	NP
35	P 3/35	NP
…	NP	NP
108	NP	P 3/108
PM	0,4175	0,6759

Table 3 : *Précision moyenne de deux requêtes ayant trois documents pertinents (notés P)et non pertinents (NP) présentés dans des rangs différents*

Pour un ensemble de requêtes, nous pouvons opter pour la moyenne arithmétique (MAP) des précisions moyennes individuelles (PM). Afin de savoir si une différence entre deux modèles s'avère statistiquement significative, nous avons opté pour un test bilatéral non-paramétrique (basée sur le ré-échantillonnage aléatoire ou *bootstrap* [SAV 97], avec un seuil de signification $\alpha = 5\ \%$). Comme nous l'avons démontré empiriquement, d'autres tests statistiques comme le *t*-test ou le test du signe aboutissent très souvent aux mêmes conclusions [SAV 06]. Dans nos tables, les différences de performance statistiquement significatives seront soulignées.

	MAP		GMAP	
	T	TD	T	TD
Okapi	0,4407	0,5058	**0,2547**	0,3644
I(n_e)C2	**0,4418**	**0,5116**	0,2474	**0,3755**
DLH	0,4076	0,4846	0,2154	0,3338
LM (λ=0,35)	0,3986	0,4721	0,2039	0,3182
Lnu-ltc	0,4066	0,4817	0,2313	0,3441
tf · idf	0,2830	0,3304	0,1129	0,1753

Table 4 : *Evaluation de nos divers modèles de dépistage selon la précision moyenne (MAP) ou la moyenne géométrique (GMAP)*

En utilisant les six modèles de recherche en fonction des requêtes très courtes (T) ou de longueur moyenne (TD), la table 4 indique les performances obtenues en recourant à la moyenne arithmétique (MAP) ou géométrique (GMAP, que nous analyserons dans la prochaine sous-section). En regardant uniquement les valeurs de la MAP, on constate que la meilleure qualité est obtenue par le modèle probabiliste I(n_e)C2. Les différences entre cette approche et les autres s'avèrent statistiquement significatives (valeurs soulignées dans la table 4) sauf avec Okapi

pour lequel la différence n'est pas statistiquement significative. L'augmentation de la longueur des requêtes de « titre seulement » (ou T) à « titre & descriptif » (ou TD) n'a pas d'influence sur le classement des divers modèles.

Cependant, l'interprétation des différences de MAP doit être faite avec précaution. Ainsi, si l'on compare les résultats des requêtes courtes (T) ou de longueur moyenne (TD) avec le modèle I(n_e)C2, on constate que l'augmentation la plus importante est obtenue avec la requête n^o 141 ("Une lettre piégée pour Arabella Kiesbauer"). En utilisant uniquement le titre, le seul document pertinent apparaît en 9^e position (PM = 0,1111). Avec la formulation TD, la précision moyenne de cette requête s'élève à 1,0 et l'unique article pertinent se place au premier rang. Certes, l'usager final constatera une différence mais, en regard des 151 requêtes, cette variation est jugée plus importante que le déplacement du premier document pertinent de la 49^e position à la première (requête n^o 54, "Résultats des demi-finales", sept documents pertinents, dont la PM passe de 0,0078 (requête T) à 0,5479 (requête TD)). Pour nous, un déplacement de la 49^e vers la première devrait avoir plus d'influence qu'un déplacement de la neuvième à la première place.

4.2. La moyenne géométrique (GMAP)

Lorsque l'on désire favoriser les systèmes proposant une qualité de réponse minimale pour toutes les interrogations, la MAP possède d'autres inconvénients. Par exemple, si une nouvelle stratégie permet d'améliorer la PM d'une requête facile de 0,5 à 0,55 (augmentation absolue de 0,05 et relative de 10 %), cette augmentation aura le même impact aux yeux de la MAP qu'une augmentation de la PM d'une requête difficile de 0,05 à 0,1 (soit + 100 %). La MAP accorde à chaque requête la même importance et ne distinguera donc pas entre ces deux cas de figures. Or, nous souhaitons justement distinguer ces deux cas en donnant plus d'influence à la seconde amélioration. Afin d'obtenir cet effet, les diverses campagnes d'évaluation (piste robuste) proposent de recourir à la moyenne géométrique (GMAP) que l'on définit selon la formule suivante.

$$\text{GMAP} = \sqrt[m]{\prod_{i=1}^{m} PM_i} = e^{1/m \sum_{i=1}^{m} \ln(PM_i)} \qquad (7)$$

dans laquelle PM_j indique la précision moyenne de la i^e requête sur les m dont on dispose. De plus, si la PM est nulle pour une requête donnée, nous la remplaçons par une très faible valeur (soit 0,0001 dans nos évaluations).

Dans la table 4 sous la colonne « GMAP », nous avons indiqué la performance de nos six modèles en fonction des requêtes T et TD. Comme on le constate, ces valeurs de performance sont fortement corrélées avec celles de la MAP (en fait le coefficient de corrélation s'élève à 0,96). L'augmentation de la longueur des requêtes de T à TD n'a pas d'influence significative sur le classement, tout au plus

une permutation des deux premiers rangs lorsque la GMAP est utilisée. Mesurer en recourant à la moyenne arithmétique ou géométrique, le classement des diverses approches reste similaire mais pas identique. En effet, le modèle Lnu possède une MAP relativement proche du modèle DLH. Par contre, aux yeux de la mesure GMAP, le modèle Lnu occupe clairement le troisième rang derrière les approches I(n_e)C2 et Okapi.

Mais la moyenne géométrique possède aussi quelques défauts. Comme pour la MAP, la valeur de cette mesure de performance est un nombre sans signification réelle. Si l'on analyse quelques requêtes, nous constatons également quelques difficultés. Ainsi, lorsque la formulation des requêtes passe de T à TD (modèle Okapi), l'accroissement le plus important selon la GMAP est obtenu pour la requête n° 200 ("Innondationeurs en Hollande et en Allemagne"). Avec le titre uniquement, aucun document pertinent n'est dépisté (PM = 0,0). Avec la requête TD, la précision moyenne s'élève à 0,0314 et le premier article pertinent se place au 43e rang. Un deuxième exemple permettra de mieux cerner les effets de cette moyenne géométrique. En reprenant le même contexte (modèle Okapi, requête T et TD), un écart très important est signalé pour la requête n° 60 ("Les affaires en France"). Avec le titre uniquement, le premier document pertinent se place en 59e position (PM = 0,0043). Avec l'interrogation TD, la première bonne réponse apparaît en première place (PM = 0,3787). Dans ce deuxième cas, l'usager final voit réellement une amélioration. Pour la moyenne géométrique, le dépistage d'un article pertinent en 43e place possède plus d'impact que le déplacement de la première bonne réponse du 59e rang en première position.

4.3. Les requêtes difficiles

Si nous regardons le rang du premier document pertinent retrouvé, on constate que, sur l'ensemble des 151 requêtes, ce rang est strictement supérieur à 20 pour douze interrogations. En posant comme limite la valeur dix, nous comptons 19 interrogations difficiles, comme l'indique la table 5. Dans cette dernière, on constate que le modèle Lnu revient sept fois comme modèle proposant le meilleur rang pour le dépistage d'un article pertinent. En particulier, ce modèle de recherche apparaît fréquemment dans le haut du tableau, c'est-à-dire face à des requêtes très difficiles. Un tel phénomène explique les bonnes valeurs GMAP de ce modèle dans la section précédente (voir table 4). De plus, nous savons que des services commerciaux ont opté pour cette stratégie vectorielle. En effet, elle présente un meilleur comportement face à des requêtes ardues, même si une telle solution s'avère significativement moins bonne que le modèle Okapi si l'on considère la MAP.

Les requêtes reprises dans la table 5 s'avèrent difficiles pour l'ensemble des stratégies de recherche. Ainsi, pour la requête n° 60, ("Les affaires en France"), le premier article pertinent dépisté par le modèle Okapi apparaît en 59e position (ou en

453e pour le modèle de langue (LM)). Par contre, pour le modèle vectoriel Lnu, le premier document pertinent se place au 15e rang.

Requête	PM	Rang	Modèle RI
no 200	0,0002	711	Lnu
no 155	0,0075	171	Lnu
no 117	0,0016	129	I(n_e)C2
no 156	0,0084	119	Okapi
no 151	0,0140	65	I(n_e)C2
no 91	0,0016	59	Lnu
no 148	0,0277	40	Lnu
no 52	0,0239	38	DLH
no 48	0,0148	37	I(n_e)C2
no 46	0,0263	36	Lnu
no 120	0,0098	32	DLH
no 135	0,0647	21	Okapi
no 51	0,3650	17	Lnu
no 60	0,0062	15	Lnu
no 109	0,0801	13	I(n_e)C2
no 113	0,0390	13	I(n_e)C2
no 111	0,0344	13	I(n_e)C2
no 177	0,0663	12	I(n_e)C2
no 182	0,0549	11	LM

Table 5 : *Liste des douze requêtes difficiles (le premier document pertinent dépisté apparaît à un rang supérieur à 10)*

Afin de connaître les raisons expliquant la difficulté sous-jacente de ces requêtes plusieurs explications peuvent être avancées [SAV 07]. Nous pouvons les résumer par la présence de fautes d'orthographe (requête no 200, "Innondationeurs en Hollande et en Allemagne"), la présence d'une liste trop longue de mots-outils (requête no 91, "AI en Amérique latine" avec "ai" forme verbale qui sera éliminé), ou la lemmatisation insuffisante ou trop radicale (requête no 117, "Elections parlementaires européennes"), une interrogation trop vague (requête no 51, "Coupe du monde de football"), une formulation qui ne permet pas de dépister les articles pertinents (requête no 52, "Dévaluation de la monnaie chinoise"), et enfin la présence de synonymes ou de particularités nationales (requête no 155, "Les risques du téléphone portable", appareil nommé "natel" en Suisse ou "cellulaire" au Québec).

4.4. Le rang du premier document pertinent

Les mesures MAP ou GMAP ne sont pas exemptes de problèmes en particulier lorsque l'on désire accorder plus d'importance aux requêtes difficiles. Comme mesure de performance alternative ou complémentaire, nous pourrions penser à la précision après 10 réponses (limite correspondant au premier écran de la liste de résultats d'un moteur de recherche). Cependant cette mesure possède le défaut de ne pas tenir compte du rang des documents pertinents, pourvu que ces derniers apparaissent dans les dix premiers. Ainsi, si l'on dépiste deux éléments pertinents et qu'on les place en première et deuxième position, la précision après 10 documents sera de 0,2. Une valeur identique s'obtient en plaçant ces deux articles à la neuvième et dixième place.

Comme autre mesure on peut recourir à la moyenne de l'inverse du rang de la première bonne réponse (MRR ou *Mean Reciprocal Rank*). Cette approche possède des avantages intéressants. Premièrement, sa valeur peut être interprétée par l'utilisateur. Deuxièmement, elle tient compte du rang, certes limité au premier document pertinent dépisté. Troisièmement, l'identification des requêtes difficiles est aisée ; elles posséderont une valeur MRR supérieure à 0,1 (soit 1/10), si l'on fixe comme critère l'absence d'article pertinent dans les dix premiers rangs. Cependant, dépister un article pertinent en première place ou en deuxième entraîne une différence très nette de la performance, soit 0,5. En effet, l'inverse du premier rang redonne la valeur 1/1 = 1 tandis qu'en deuxième position, cette performance sera de 1/2 = 0,5.

	FRS	
	T	TD
Okapi	**0,8221**	0,8984
I(n_e)C2	0,8112	**0,9019**
DLH	0,7968	0,8767
LM (λ=0,35)	0,7743	0,8680
Lnu-ltc	0,8061	0,8932

Table 6 : *Evaluation de nos divers modèles de dépistage selon l'inverse pondéré du rang du premier article pertinent*

Afin de réduire l'importance accordée à la première place, on peut recourir au score pondéré du premier document pertinent (FRS ou *First Relevant Score*) défini comme $K^{(1-r)}$, avec r le rang de la première bonne réponse et K une constante (fixée à 1,08 dans notre étude) [TOM 06]. Si nous rencontrons la première bonne réponse en première place, le score sera de 1, comme dans la mesure MRR. Ensuite, pour le deuxième rang, nous obtenons la valeur de 0,926 (au lieu de 0,5), pour le troisième rang la valeur 0,857 (au lieu de 0,333) et 0,794 (au lieu de 0,25) pour le quatrième.

Ce score décroît de manière exponentielle pour atteindre 0,5 au dixième rang. Le rang pour lequel le score obtient la valeur 0,5 détermine la constante *K*, (soit 1,08 dans notre cas pour obtenir la valeur 0,5 pour la dixième position). La différence entre la première et la deuxième place s'atténue par rapport à la mesure MRR et correspond mieux, à nos yeux, à l'appréciation de l'utilisateur. Enfin, si aucun article pertinent n'est dépisté, la valeur de *r* est fixée arbitrairement à 1001.

Dans la table 6, nous avons repris cette mesure FRS avec nos deux types de requêtes et nos différentes stratégies de recherche. Les deux classements correspondent exactement à ceux obtenus avec la moyenne géométrique (GMAP, voir table 4). La mesure FRS accorde donc plus de poids aux requêtes difficiles (pour lesquelles le rang du premier document pertinent sera plus élevé). Nous pouvons compléter cette analyse par un test statistique basé sur la technique du ré-échantillonnage aléatoire (*bootstrap*) [SAV 97] (les différences significatives par rapport à la performance la plus élevée sont soulignées dans la table 6). Toutefois, les mesures GMAP et FRS n'aboutissent pas toujours à des résultats identiques comme le démontre l'analyse présentée dans la prochaine section.

5. Application à l'expansion aveugle des requêtes

Afin de vérifier la cohérence des conclusions que l'on peut déduire avec les diverses mesures d'évaluation présentées dans la section précédente, nous avons choisi d'analyser l'expansion automatique des requêtes [ROC 71], [EFI 06]. Plusieurs études indiquent que le recours à cette technique voire à l'expansion aveugle de la requête [BUC 96] permet d'améliorer significativement la performance moyenne. Nous avons appliqué une telle stratégie sur les deux modèles proposant la meilleure performance soit le modèle Okapi et I(n_e)C2.

Néanmoins, une discussion préliminaire s'impose. En effet, si l'on désire améliorer la qualité de la réponse et, en particulier, pour les requêtes difficiles, l'expansion aveugle de la requête n'a aucune chance d'atteindre cet objectif. En effet, une requête ardue se définit comme une liste de résultats sans aucun article pertinent parmi les premières dix références retrouvées. Or l'expansion aveugle s'appuie justement sur cet ensemble pour y extraire de nouveaux termes. Si une telle remarque relève du bon sens, la réalité dévoile une autre facette. Ainsi, par exemple pour la requête n° 46 ("Embargo sur l'Iraq"), le premier document pertinent apparaît en position 55 avec le modèle Okapi. Après expansion (3 documents / 20 termes), la quatrième place est occupée par le premier article pertinent.

La table 7 indique les trois mesures de performance pour le modèle Okapi et la table 8 pour le modèle I(n_e)C2. Dans les deux cas, la même stratégie d'expansion aveugle (Rocchio [BUC 96]) avec les mêmes paramètres a été appliquée. Au regard de la MAP et de la GMAP, cette expansion permet d'améliorer significativement la

performance pour le modèle Okapi tandis qu'avec le modèle I(n_e)C2, la qualité est statistiquement inférieure après l'expansion automatique.

	MAP	GMAP	FRS
Modèle avant	0,4407	0,2547	**0,8221**
3 doc / 20 termes	**0,4873**	**0,2759**	0,7819
5 doc / 20 termes	0,4751	0,2697	0,7724
10 doc / 20 terms	0,4815	0,2743	0,7728

Table 7 : *Evaluation avant et après l'expansion automatique de la requête modèle Okapi, requête « titre » seulement*

	MAP	GMAP	FRS
Modèle avant	**0,4418**	**0,2474**	**0,8112**
3 doc / 20 termes	0,4027	0,1865	0,7279
5 doc / 20 termes	0,4041	0,1883	0,7220
10 doc / 20 terms	0,3791	0,1986	0,7067

Table 8 : *Evaluation avant et après l'expansion automatique de la requête modèle I(n_e)C2, requête « titre » seulement*

Pour le modèle Okapi (table 7), cette amélioration permet de faire passer la MAP de 0,4407 jusqu'à 0,4873, un accroissement relatif de 10,5% (ou de 8,3 % avec la GMAP). Un test statistique indique que cette modification s'avère significative. Une inspection requête par requête montre que cette stratégie améliore la précision moyenne dans 90 cas, la dégrade dans 46 cas et pour les 15 requêtes restantes, la précision moyenne demeure inchangée. Avec le modèle I(n_e)C2 (table 8), ces résultats ne se confirment pas. L'expansion aveugle entraîne une diminution de la précision moyenne pour 81 interrogations, l'améliore dans 61 cas (pour 9 requêtes, la précision moyenne reste la même).

Si l'on reprend cette analyse au regard de l'inverse pondéré du rang du premier document pertinent dépisté (voir les colonnes FRS dans les table 7 et 8), la technique de l'expansion automatique des requêtes ne s'avère plus aussi attractive. Sur les 151 requêtes et avec le modèle Okapi, on ne constate aucun changement pour 93 interrogations ; le rang du premier document pertinent dépisté reste inchangé. Pour 31 requêtes, ce rang s'accroît après l'expansion de requêtes. Dans ces cas, l'expansion produit un effet négatif sur la liste des résultats. Enfin, pour 27 requêtes l'expansion automatique génère un effet positif en déplaçant plus près du sommet de la liste un document pertinent. Aux yeux de ce critère de performance, la stratégie d'expansion aveugle génère une détérioration significative de la performance. Le rang du premier article pertinent augmente. Toutefois, cette mesure se base exclusivement sur le rang du premier document pertinent et donc ne

tient pas compte du rappel. Nous pensons donc qu'une telle mesure doit être vue comme complémentaire à une évaluation basée sur la moyenne arithmétique ou géométrique des précisions moyennes.

6. Conclusion

Sur la base d'un corpus d'articles de journaux rédigés en langue française et de 151 requêtes, nous avons démontré que le modèle Okapi ou une approche dérivée du paradigme *Divergence from Randomness* apporte la meilleure performance. Cependant, l'interprétation de la moyenne arithmétique des précisions moyennes (MAP) et des différences entres approches doit être faite avec précaution. Nous avons illustré par quelques exemples les difficultés sous-jacentes à la lecture et à toute comparaison utilisant la précision moyenne ou la MAP.

Cette communication a également abordé le problème de l'évaluation robuste accordant une importance plus grande aux requêtes difficiles. En recourant à la moyenne géométrique (GMAP), nous avons démontré que cette mesure place sous un meilleur jour les performances obtenues par le modèle vectoriel Lnu.

Comme la moyenne géométrique n'est pas exempt de reproches, nous proposons de recourir à une mesure complémentaire, soit la FRS [TOM 06] valeur basée sur l'inverse pondéré du rang du premier document pertinent. En analysant l'expansion aveugle des requêtes [BUC 96], nous avons obtenu des résultats quelque peu contradictoires. D'une part, cette stratégie améliore la performance mesurée par la MAP et seulement dans le cas du modèle Okapi. Par contre, en analysant la qualité de la réponse en considérant le rang du premier document dépisté, cette stratégie détériore significativement les performances pour les deux modèles de recherche étudiés. Ces deux mesures mettent en lumière des phénomènes distincts et devraient s'utiliser conjointement afin d'obtenir une meilleure appréciation de la performance ou de la différence de performances entre deux stratégies de recherche.

Remerciements

Cette recherche a été financée en partie par le Fonds national suisse pour la recherche scientifique (subsides n^0 200020-103420 et n^0 200020-115866).

7. Bibliographie

[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, n° 4, 2002, p. 357-389.

[AMA 06] Amati, G. "Frequentist and Bayesian approach to information retrieval", Proceedings ECIR 2006, LNCS #3936, Springer, Berlin, 2006, p. 13-24.

[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.

[CAR 05] Carmel, D., Yom-Tov, E., & Soboroff, I. "Predicting query difficulty – Methods and applications", ACM-SIGIR Forum, vol. 39, n° 2, 2005, p. 25-28.

[CAR 06] Carmel, D., Yom-Tov, E., Darlow, A. & Pelleg, D. "What makes a query difficult?", Proceedings of ACM-SIGIR'2006, 2006, p. 390-397.

[EFI 96] Efthimiadis, E.N. "Query expansion", Annual Review of Information Science and Technology, 31, 1996, p. 121-187.

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

[PET 02] Peters, C., Braschler, M., Gonzalo, J., & Kluck, M. (Eds). "Evaluation of cross-language information retrieval", LNCS #2406, Springer, Berlin, 2002.

[PET 03] Peters, C., Braschler, M., Gonzalo, J., & Kluck, M. (Eds). "Advances in cross-language information retrieval", LNCS #2785, Springer, Berlin, 2003.

[PET 04] Peters, C., Braschler, M., Gonzalo, J., & Kluck, M. (Eds). "Comparative evaluation of multilingual information access systems", LNCS #3237, Springer, Berlin, 2004.

[PLA 05] Plachouras, V., He, B., & Ounis, I. "University of Glasgow at TREC2004: Experiments in web, robust and terabytes tracks with Terrier", Proceedings of TREC-2005, NIST Publication #500-261, Gaithersburg (MD), 2005.

[ROB 00] Robertson, S.E., Walker, S., & Beaulieu, M. "Experimentation as a way of life: Okapi at TREC", Information Processing & Management, vol. 36, n° 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 (NJ), 1971, p. 313-323

[SAV 97] Savoy, J. "Statistical inference in retrieval effectiveness evaluation", Information Processing & Management, vol. 33, n° 4, 1997, p. 495-512.

[SAV 02] Savoy, J. "Recherche d'informations dans des corpus en langue française : Utilisation du référentiel Amarylis", TSI, Technique et Science Informatiques, vol. 21, n° 3, 2002, p. 345-373.

[SAV 06] Savoy, J. "Un regard statistique sur l'évaluation de performance : L'exemple de CLEF 2005", Actes 3ième COnférence en Recherche d'Information et Applications CORIA'06, Lyon, 2006, p. 73-84.

[SAV 07] Savoy, J. "Why do successful search systems fail for some topics", Proceedings ACM-SAC, The ACM Press, 2007, to appear.

[TOM 06] Tomlinson, S. "Bulgarian and Hungarian experiments with Hummingbird™ SearchServer at CLEF 2005", In Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., & de Rijke, M. (Eds). "Accessing multilingual information repositories", LNCS #4022, Springer, Berlin, 2006, p. 194-203.

[VOO 05] Voorhees, E.M. "Overview of the TREC 2004 robust retrieval track", Proceedings of TREC-2004, NIST Publication#500-261, Gaithersburg (MD), 2005.

[VOO 06] Voorhees, E.M. "The TREC 2005 robust track", ACM-SIGIR Forum, vol. 40, n° 1, 2006, p. 41-48.