Comparaison des stratégies d'indexation pour les langues asiatiques

Samir Abdou

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

RÉSUMÉ. *En recherche d'information, les langues chinoise et japonais présentent des défis multiples. Contrairement aux langues européennes, les mots ne se sont pas délimités de manière explicite ce qui pose un problème pour l'indexation. Pour cette raison, plusieurs travaux ont proposé différentes stratégies pour représenter les documents (et requêtes) rédigés dans ces langues. Cet article présente une comparaison des stratégies d'indexation les plus courantes. En particulier, nous avons comparé quatre stratégies pour le chinois (unigrammes, bigrammes, uni- et bigrammes et finalement les mots), deux pour le japonais (bigrammes et mots) et trois pour le coréen (mots, bigrammes et morphèmes). Basé sur les collections-tests de NTCIR-5, nous avons évalués ces différentes approches à l'aide de neuf modèles de recherche, soit deux approches probabilistes et sept vectoriels.*

ABSTRACT. *In information retrieval, Chinese and Japanese present many challenging problems. Unlike most European languages, the lack of explicit word boundaries represents one of the most important issue for indexing. For this reason, many works proposed different approaches to index documents or requests written in these languages. This article presents a comparison of the common indexing strategies. More precisely, we compared four strategies for Chinese (1-grams, 2-grams, 1- & 2-grams and words), two for Japanese (2-grams, and words) and three for Korean (words, 2-grams, and morphemes). Using test collections of the NTCIR-5, we evaluated these various approaches on nine retrieval schemes: two probabilistic models and seven vectorial models.*

MOTS-CLÉS : *Recherche monolingue ; stratégies d'indexation ; langues chinoise, japonaise et coréenne.*

KEYWORDS: *Monolingual search; indexing strategies; Chinese, Japanese and Korean languages.*

1. Introduction

Dans le domaine de la recherche d'information (RI), l'évolution très rapide d'Internet permettant l'émergence d'un savoir planétaire partagé génère de multiples défis. Outre le volume d'informations toujours plus important mis à disposition ou la présence de multiples supports, le caractère plurilingue de la Toile représente à nos yeux un enjeu considérable. Dans ce contexte, l'importance grandissante d'autres langues 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. En septembre 2005[1], la proportion d'internautes naviguant en anglais était estimée à 31,7 % contre 13 % pour le chinois, 8,1 % pour le japonais et 3,4 % pour le coréen. Sur cette base, on estime que l'utilisation des langues asiatiques sur le Web et en particulier du chinois va atteindre des valeurs comparables ou supérieures à celle de l'anglais.

En comparaison de l'anglais ou d'autres langues indo-européennes, les langues chinoise, japonaise ou coréenne présentent des caractéristiques singulières. Premièrement, les mots ne sont plus explicitement marqués en chinois ou en japonais. Dans ces deux langues, une phrase est une succession de symboles sans espaces qui se termine par une ponctuation claire. Lors de l'indexation d'un document, une étape préliminaire vise à segmenter les phrases et cette segmentation, habituellement automatique, peut s'opérer de plusieurs manières. On peut recourir à une subdivision en blocs de taille fixe ou selon des informations lexicales ou statistiques. Deuxièmement, le nombre d'idéogrammes s'avère très élevé (par exemple plus de 13 000 pour le chinois traditionnel) comparé à nos 26 lettres. La langue japonaise combine, à cette écriture chinoise, 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, les mots sont explicitement délimités mais, à l'exemple de l'allemand, cette langue possède de très nombreux mots composés, habituellement générés par concaténation et adjonction de divers mots simples et suffixes.

Face à ces défis et sous l'impulsion des campagnes d'évaluation TREC[2] et NTCIR, diverses approches ont été développées pour indexer et extraire des documents ou réponses rédigés dans l'une de ces langues asiatiques. Dans le cadre de cet article, nous souhaitons comparer la performance de différentes stratégies de segmentation pour dépister efficacement des documents écrits en chinois, japonais ou coréen. La plupart des études antérieures évaluent les différentes approches possibles à l'aide d'un nombre limité de moteurs de recherche. Etant donnée le nombre de paramètres et de processus sous-jacents (requêtes structurés, expansion de requêtes, combinaisons diverses), la comparaison directe ne permet souvent pas de connaître l'influence de ces diverses composantes au regard de stratégies de recherche différentes.

[1] Voir le site http://www.internetworldstats.com
[2] Voir les sites : http://trec.nist.gov et http://research.nii.ac.jp/ntcir

La suite de cet article est organisée de la manière suivante. La deuxième section présentera quelques résultats de travaux antérieurs concernant les méthodes d'indexation. Dédiée à notre méthodologie d'évaluation, la troisième section décrira les collections-tests sur lesquelles nous avons effectué nos expériences d'une part et, d'autre part, les stratégies d'indexation et de recherche utilisées. Dans la quatrième section, nous aborderons l'évaluation de diverses représentations et stratégies de recherche pour les trois langues asiatiques retenues. Enfin, les principaux résultats de cette étude seront synthétisés dans la cinquième section.

2. Etat des connaissances

Afin de représenter des documents rédigés en langue chinoise, japonaise ou coréenne, on distingue généralement trois approches différentes. Premièrement, les méthodes peuvent se baser sur des séquences de caractères en découpant le texte en *n*-grammes. Ce choix présente l'avantage de ne pas nécessiter de connaissances linguistiques 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, 2002 ; Peters, 2005) mais requiert un temps de réponse plus long (environ 10 fois plus important). Deuxièmement, on peut recourir à des connaissances linguistiques ou s'appuyer sur une étude statistique des corpus pour reconnaître et extraire les mots d'un texte. Cependant, la non exhaustivité 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 de diverses manières.

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 par 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, TREC-9 et NTCIR-2), 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 (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 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 bigramme s'est avérée similaire à une approche par trigramme (McNamee, 2001).

Finalement, pour la langue coréenne, Lee & Ahn (1996) ont proposé une approche par bigrammes obtenues 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. 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.

3. Méthodologie

3.1. *Les collections-tests*

Les collections tests utilisées dans nos expériences sont issues de la cinquième campagne d'évaluation de NTCIR (Kishida *et al.*, 2005). En comparant ces corpus (tableau 1), on constate que la taille et le nombre de documents sont similaires pour le chinois et le japonais tandis que le coréen représente un corpus plus restreint.

	Chinois	Japonais	Coréen
Taille (en Mo)	1 100	1 100	312
nombre de documents	901 446	854 400	220 374
codage	BIG5	EUC_JP	EUC_KR
nombre de requêtes	50	47	50
nb docs pertinents / requête	37,7	44,94	36,58

Tableau 1. *Statistiques sur les collections-tests*

Dans ces collections-tests, les jugements de pertinence sont définis selon quatre degrés, à savoir : « très pertinent », « pertinent », « partiellement pertinent » et « non pertinent ». En suivant les règles appliquées lors de l'évaluation officielle, nous considérons dans cet article comme « pertinent » que les documents jugés « très pertinent » et « pertinent » (*rigid evaluation*).

Comme besoins d'information, nous disposons de 50 requêtes disponibles dans chacune des trois langues. Dans le tableau 1, on constate cependant que le corpus japonais n'offre que 47 requêtes car pour trois demandes, aucun document pertinent n'a été trouvé dans ce corpus. En comparant la moyenne des documents pertinents

par requête, le corpus japonais présente une valeur nettement supérieure (44,94) à celle du corpus coréen (36,58) ou chinois (37,7). Suivant le format traditionnel de TREC, chaque requête est composée de cinq champs, à savoir un identificateur, un titre, une section descriptive, une section narrative et finalement une section de concepts liés la requête. Dans les différentes expériences présentées dans cet article, nous avons utilisé uniquement les requêtes construites à l'aide de la section « titre ».

3.2. *Stratégies d'indexation*

Afin d'indexer les documents et les requêtes, nous avons adopté différentes stratégies d'indexation. Premièrement, nous avons utilisé une méthode basée simplement sur les caractères ou unigrammes. Dans ce cas, chaque caractère ou idéogramme d'une séquence donnée représente une unité d'indexation distincte. Nous avons appliqué cette approche uniquement pour la langue chinoise. Deuxièmement, nous avons indexé les collections disponibles dans les trois langues en recourant aux bigrammes. Dans ce cas, on génère tous les couples ordonnées de deux caractères successifs à partir d'une séquence donnée. Selon cette méthode de segmentation, la chaîne « ABCD EFG » produira l'ensemble des bigrammes {AB BC CD EF FG}. Comme alternative et pour le chinois uniquement, nous avons combiné les méthodes unigrammes et bigrammes. Ainsi sur la base de la séquence précédente, cette approche produira les unités d'indexation {A B C D E F G AB BC CD EF FG}. Troisièmement, nous avons également évalué l'indexation basée sur les mots. Afin de les définir à partir d'une phrase donnée, nous avons eu recours à des outils de segmentation automatique, à savoir *MTSeg*[3] pour la langue chinoise et *ChaSen* (Matsumoto *et al.*, 1999) pour le japonais. Pour la langue coréenne, nous avons opté pour deux représentations possibles. La première indexe les mots sans aucun traitement morphologique. La seconde s'appuie sur les morphèmes obtenus à l'aide de l'analyseur morphologique *HAM*[4] qui opère, de manière automatique, une décomposition des mots et procède à l'élimination des suffixes.

Lors de la génération des *n*-grammes, les espaces, chiffres, caractères latins ainsi que les divers signes de ponctuation constituent des marqueurs délimitant la fin de la séquence. Ces marqueurs à l'exception des caractères latins sont ensuite ignorés. Les mots écrits avec notre alphabet ne sont pas fragmentés mais utilisés tels quels. Pour le japonais, nous avons exploité le changement de classe de caractères pour séparer la génération des bigrammes. De plus, les caractères *hiragana* sont simplement éliminés car ces derniers servent essentiellement à écrire les mots-outils et les inflexions morphologiques (liées aux divers cas grammaticaux). Dans la langue nippone, les bigrammes ne comprendront donc soit uniquement des *kanji* (écriture chinoise) ou des *katakana*.

Dans l'espoir de réduire le bruit lors de la recherche, nous avons défini une liste de termes fréquents et pas ou peu porteurs d'information. Ceux-ci seront éliminés.

[3] *MandarinTools Segmenter* : http://www.mandarintools.com/segmenter.html
[4] *Hangul Analyser Module* : http://nlp.kookmin.ac.kr

Pour la langue chinoise, nous avons construit une liste comprenant 90 unigrammes, une liste composée de 49 bigrammes et, enfin, une liste de 91 mots. Pour la langue japonaise, nous avons déterminé une liste de 20 bigrammes et une liste de 30 mots. Enfin pour la langue coréenne, nous avons construit une liste de 91 bigrammes et une liste de 89 mots.

3.3. *Stratégies de recherche*

Afin de déterminer le degré de similarité entre le document D_i et la requête Q, notre système calcule le produit interne définit comme :

$$SIM(D_i, Q) = \sum_{j=1}^{m} w_{ij} \cdot w_{qj} \quad (1)$$

avec *m* indiquant le nombre de termes en communs entre le document D_i et la requête Q, w_{ij} représente le poids du terme T_j dans le document D_i et w_{qj} correspond au poids du même terme dans la requête. La pondération de ces termes peut suivre les différentes possibilités présentées dans le tableau 2.

nnn	$w_{ij} = tf_{ij}$	ltn	$w_{ij} = (\ln(tf_{ij}) + 1) \cdot idf_j$
dtn	$w_{ij} = (\ln(\ln(tf_{ij}) + 1) + 1) \cdot idf_j$	ntc	$w_{ij} = tf_{ij} \cdot idf_j \Big/ \sqrt{\sum_{k=1}^{t} (tf_{ik} \cdot idf_k)^2}$
npn	$w_{ij} = tf_{ij} \cdot \ln\left(n - df_j \Big/ df_j \right)$	atn	$w_{ij} = \left(0{,}5 + 0{,}5 \cdot (tf_{ij} \Big/ \max tf_{ij}) \right) \cdot idf_j$
dtu	$w_{ij} = \frac{(\ln(\ln(tf_{ij}) + 1) + 1) \cdot idf_j}{(1 - slope) \cdot pivot + slope \cdot nt_i}$	Lnu	$w_{ij} = \frac{(\ln(tf_{ij}) + 1) \big/ (\ln(mean\ tf_i) + 1)}{(1 - slope) \cdot pivot + slope \cdot nt_i}$
lnc	$w_{ij} = \frac{(\ln(tf_{ij}) + 1)}{\sqrt{\sum_{k=1}^{t} (\ln(tf_{ik}) + 1)^2}}$	lnc	$w_{ij} = \frac{(\ln(tf_{ij}) + 1) \cdot idf_j}{\sqrt{\sum_{k=1}^{t} (\ln(tf_{ik}) + 1) \cdot idf_k)^2}}$
Okapi	$w_{ij} = \frac{(k_1 + 1) \cdot tf_{ij}}{K + tf_{ij}}$, avec $K = k_1 \cdot \left[(1 - b) + b \cdot \frac{length_i}{avdl} \right]$		

Tableau 2. *Nos différentes formules de pondération*

Pour décrire un modèle de recherche, nous utilisons une paire de lettres indiquant la pondération utilisée pour les documents suivie de celle appliquée aux requêtes. Par exemple, le modèle « Lnu-ltc » indique que les termes dans les documents sont pondérés suivant le schéma « Lnu » tandis la pondération « ltc » sera appliquée aux termes de la requête.

En plus des modèles vectoriels, nous avons également évalué deux modèles probabilistes, à savoir le modèle Okapi (Robertson *et al.*, 2000) et le modèle PB2 issu de la famille *Divergence from randomness* (Amati & van Rijsbergen, 2002). Dans ce dernier cas, la pondération w_{ij} du terme T_j dans le document D_i combine deux mesures d'informations suivant la formule :

$$w_{ij} = Inf_{ij}^1 \cdot Inf_{ij}^2 = (-\log_2[(e^{\lambda_j} \cdot \lambda^{tfn_{ij}}) / tfn_{ij}!]) \cdot (1 - Prob_{ij}^2) \quad (2)$$

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

$$\text{avec } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((c \cdot avdl)/l_i)] \text{ et } \lambda = tc_j / n$$

dans laquelle l_i représente la longueur du document D_i, df_j indique le nombre de documents dans lesquels le terme T_j apparaît, tc_j la fréquence du terme T_j dans la collection et *n* le nombre de document dans le corpus. Dans nos évaluations, les valeurs des constantes *b*, k_1, *c* et *slope*, fixées empiriquement sur la base des collections-tests de NTCIR-4, sont données en annexe. Les constantes *avdl* et *pivot* sont calculées sur la base des corpus de NTCIR-5.

4. Evaluation

Afin de mesurer la performance d'un système de recherche, nous avons choisi la précision moyenne calculée par le logiciel `trec_eval`. Pour vérifier si la différence de performance entre deux systèmes est statistiquement significative, nous avons adopté le test du signe avec un niveau de signification de 5 % (test bilatéral). Dans l'application de ce test, l'hypothèse nulle (notée H_0) admet que les performances des moteurs sont identiques. Toutefois, si le test statistique indique que l'on doit accepter H_0, cette décision ne signifie pas que cette hypothèse H_0 soit vraie mais que, sur la base des résultats obtenus, nous ne disposons pas de suffisamment d'évidence pour contredire H_0 en faveur de l'hypothèse alternative.

Dans la présentation des résultats de nos évaluations, nous cherchons à répondre à deux questions, à savoir quelles sont les stratégies de recherche offrant les meilleures performances d'une part et, d'autre part, quelles sont les représentations les plus efficaces. Sur la base de nos tableaux d'évaluation (voir par exemple, le tableau 3), nous comparerons les lignes entre elles pour déterminer la meilleure stratégie de recherche. Pour connaître la meilleure forme d'indexation, nous comparerons les colonnes entre elles. Dans la présentation de nos évaluations, nous avons noté en gras la meilleure performance d'une colonne et les différences statistiquement significatives par rapport à cette performance seront indiquées par un astérisque (« * »). Finalement, pour comparer l'efficacité des différentes formes d'indexation, les pourcentages de différences seront calculés par rapport à la deuxième colonne et celles qui sont statistiquement significatives seront soulignées.

L'évaluation des diverses stratégies d'indexation et de recherche retenues est présentée dans le tableau 3 pour la langue japonaise. La meilleure performance est obtenue par le modèle probabiliste PB2, que l'on considère l'indexation par bigrammes ou par mots. Si l'on analyse l'indexation par bigrammes (deuxième colonne), la précision moyenne du modèle PB2 (0,2816) est suivie par les approches Okapi (0,2660) ou « ltn-ntc » (0,2651). D'un point de vue statistique, seul l'écart observé avec les modèles « ltn-ntc » et « Lnu-ltc » s'avère significatif. Si l'on compare les performances obtenues avec l'indexation par mots, la différence entre les modèles PB2 (0,3063) et Okapi (0,2655, différence relative de 15,4 %) s'avère

statistiquement significative. Avec cette forme d'indexation, les différences entre PB2 et les modèles « Lnu-ltc » (0,2743), « ltn-ntc » (0,2723) et « dtu-dtn » (0,2735) ne sont pas statistiquement significatives, contrairement aux autres approches reprises sous cette colonne.

	Précision moyenne (% changement)	
Modèle	bigramme référence	mot
PB2-nnn	**0,2816**	**0,3063** (+ 8,8 %)
Okapi-npn	0,2660*	0,2655* (- 0,2 %)
Lnu-ltc	0,2579	0,2743 (+ 6,4 %)
ltn-ntc	0,2651	0,2723 (+ 2,7 %)
dtu-dtn	0,2461*	0,2735 (+ 11,1 %)
atn-ntc	0,1799*	0,2109* (+ 17,2 %)
ntc-ntc	0,1292*	0,1227* (- 5,0 %)
ltc-ltc	0,0992*	0,0945* (- 4,7 %)
lnc-ltc	0,1070*	0,1132* (+ 5,8 %)
moyenne		+ 4,7 %

Tableau 3. *Performances des différentes moteurs de recherche (corpus en langue japonaise, 47 requêtes)*

Si l'on compare les deux stratégies d'indexation pour la langue japonaise (bigramme dans la deuxième colonne du tableau 3, mot dans la troisième colonne), nous remarquons que notre test statistique ne distingue pas de différence statistiquement significative (aucune valeur n'est soulignée). En moyenne pour les neuf stratégies de dépistage, l'indexation par mots apporte une précision moyenne supérieure de l'ordre de 4,7 % par rapport à une approche basée sur les bigrammes.

	Précision moyenne (% changement)	
Modèle	unigramme référence	bigramme
PB2-nnn	0,2774	**0,3042** (+ 9,7 %)
Okapi-npn	0,2879	0,2995 (+ 4,0 %)
Lnu-ltc	**0,2883**	0,2999 (+ 4,0 %)
ltn-ntc	0,2348*	0,2886 (+ 22,9 %)
dtu-dtn	0,2743*	0,2867 (+ 4,5 %)
atn-ntc	0,2329*	0,2527* (+ 8,5 %)
ntc-ntc	0,1162*	0,2130* (+ 83,3 %)
ltc-ltc	0,1464*	0,1933* (+ 32,0 %)
lnc-ltc	0,1992*	0,2053* (+ 3,1 %)
moyenne		+ 19,1 %

Tableau 4a. *Performances des différents moteurs de recherche (corpus en langue chinoise, 50 requêtes)*

Modèle	Précision moyenne (% changement)			
	bigramme référence	mot	unigram. & bigram.	uni. & bi. vs. mot
PB2-nnn	**0,3042**	**0,3246** (+ 6,7 %)	**0,3433** (+ 12,9 %)	+ 5,4 %
Okapi-npn	0,2995	0,3230 (+ 7,8 %)	0,3321 (+ 10,9 %)	+ 2,7 %
Lnu-ltc	0,2999	0,3227 (+ 7,6 %)	0,3355 (+ 11,9 %)	+ 3,8 %
ltn-ntc	0,2886	0,2833 (- 1,8 %)	0,3068 (+ 6,3 %)	+ 7,7 %
dtu-dtn	0,2867	0,2894 (+ 0,9 %)	0,3094 (+ 7,9 %)	+ 6,5 %
atn-ntc	0,2527*	0,2578* (+ 2,0 %)	0,2729* (+ 8,0 %)	+ 5,5 %
ntc-ntc	0,2130*	0,1645* (- 22,8 %)	0,2201* (+ 3,3 %)	+ 25,3 %
ltc-ltc	0,1933*	0,1772* (- 8,3 %)	0,2202* (+ 13,9 %)	+ 19,5 %
lnc-ltc	0,2053*	0,2189* (+ 6,6 %)	0,2309* (+ 12,5 %)	+ 5,2 %
moyenne		- 0,1 %	+ 9,7 %	+ 10,8 %

Tableau 4b. *Performances des différents moteurs de recherche (corpus en langue chinoise, 50 requêtes)*

En analysant les tableaux 4a et 4b (évaluation en langue chinoise), nous constatons qu'à l'exception de l'indexation par unigrammes (deuxième colonne du tableau 4a), le modèle probabiliste PB2 offre toujours la meilleure performance. Toutefois, la différence entre PB2 et les modèles Okapi et « Lnu-ltc » n'est jamais significative. Présentant aussi une performance assez élevée et dont parfois la différence avec la meilleure approche s'avère significative, on retrouve les modèles vectoriels « dtu-dtn » et « ltn-ntc ».

En comparant les stratégies d'indexation unigrammes et bigrammes (tableau 4a), nous constatons que cette dernière permet systématiquement une précision moyenne supérieure à la première. Comme l'indique la dernière ligne du tableau 4a, la différence relative s'élève, en moyenne, à 19,1 % par rapport à une indexation par unigrammes. Par contre, les différences de performance, modèle par modèle, ne sont pas significatives, sauf pour les approches « ltn-ntc », « ntc-ntc » et « ltc-ltc ».

Fixons comme référence les performances obtenues par l'indexation par bigrammes. Nous pouvons la comparer à une indexation par mots (troisième colonne du tableau 4b) ou avec une indexation combinée "unigramme & bigramme" (quatrième colonne). Des précisions moyennes indiquées dans le tableau 4b, on constate que dans l'ensemble l'indexation combinée offre une meilleure performance avec, en moyenne, une amélioration de 9,7 % comparée à l'indexation par bigrammes. En comparant les modèles les uns avec les autres, les différences de performance sont statistiquement significatives et en faveur de l'indexation combinée, à l'exception du modèle « ntc-ntc ». Finalement, en comparant l'indexation combinée et l'indexation par mots (dernière colonne), nous remarquons que si l'indexation combinée offre une précision moyenne supérieure, les deux formes d'indexation proposent une performance jugée statistiquement similaire.

Pour le corpus coréen (tableau 5), nous constatons que le modèle présentant la meilleure performance dépend de la stratégie d'indexation adoptée. Ainsi avec une

indexation par mots simples (sans aucun traitement morphologique), le modèle « dtu-dtn » propose la meilleure précision moyenne. Toutefois, cette performance ne s'écarte pas, statistiquement, de la deuxième meilleure performance obtenue par le modèle PB2 (0,2378) ou « ltn-ntc » (0,2370). Lorsque l'on adopte une indexation par morphèmes (troisième colonne), la meilleure performance est obtenue avec le modèle PB2 suivi des modèles « Lnu-ltc » (0,3560) et Okapi (0,3549), sans que les différences observées entre ces approches soient statistiquement significatives. Finalement, pour l'indexation par bigrammes, nous constatons que le modèle vectoriel « Lnu-ltc » (0,3973) offre la meilleure performance. Dans ce cas, les différences observées avec les deux modèles probabilistes ou avec « dtu-dtn » ne sont pas jugées statistiquement significatives.

	Précision moyenne (% changement)			
Modèle	mot simple référence	morphème (HAM)	bigramme	bigram vs. morphème
PB2-nnn	0,2378	**0,3659** (+ 59,9 %)	0,3729 (+ 56,8 %)	+ 1,9 %
Okapi-npn	0,2245*	0,3549 (+ 58,1 %)	0,3630 (+ 61,7 %)	+ 2,3 %
Lnu-ltc	0,2296	0,3560 (+ 55,1 %)	**0,3973** (+ 73,0 %)	+ 11,6 %
ltn-ntc	0,2370	0,3383 (+ 42,7 %)	0,3708 (+ 56,5 %)	+ 9,6 %
dtu-dtn	**0,2411**	0,3339* (+ 38,5 %)	0,3673 (+ 52,3 %)	+ 10,0 %
atn-ntc	0,2242*	0,2983* (+ 33,1 %)	0,3270* (+ 45,9 %)	+ 9,6 %
ntc-ntc	0,1548*	0,2324* (+ 50,1 %)	0,2506* (+ 61,9 %)	+ 7,8 %
ltc-ltc	0,1606*	0,2299* (+ 43,2 %)	0,2260* (+ 40,7 %)	- 1,7 %
lnc-ltc	0,1452*	0,2322* (+ 59,9 %)	0,2414* (+ 66,3 %)	+ 4,0 %
moyenne		+ 48,3 %	+ 57,2 %	+ 6,1 %

Tableau 5. *Performances des différents moteurs de recherche (corpus en langue coréenne, 50 requêtes)*

Si l'on compare les trois stratégies d'indexation pour le corpus coréen, nous remarquons que l'indexation par mots simples (deuxième colonne) propose une performance moindre. En moyenne et sur les neuf moteurs étudiés, la différence relative est de 48 % (morphèmes) ou de 57 % (bigrammes). De plus, quelque soit le modèle considéré, les différences de performance sont statistiquement significatives (et donc soulignées). En revanche, lorsqu'on compare l'indexation par morphèmes à celle par bigrammes (variation relative indiquée dans la dernière colonne), on remarque une performance plus élevée avec l'indexation par bigrammes. Par contre, cette différence peut être peu élevé (1,9 % au regard du modèle PB2) et souvent non significative (sauf pour deux modèles, soit « Lnu-ltc » et « ltn-ntc »).

5. Conclusion

Basé sur 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 principales conclusions que nous pouvons en tirer sont les suivantes.

Le modèle probabiliste PB2 issu de la famille *Divergence from randomness* (Amati & van Rijsbergen, 2002) offre souvent la performance la plus élevée, indépendamment de la langue ou de la forme d'indexation (*n*-gramme ou mot). Comme alternative, nous pouvons suggérer le modèle Okapi ou les approches vectorielles « Lnu-ltc » ou « dtu-dtn ». En effet, les différences de performance entre ces systèmes de dépistage n'est souvent pas statistiquement significative, indiquant que la modification de quelques requêtes ou la modification du corpus peut modifier les classements présentés.

Pour les langues japonaise (tableau 3) et chinoise (tableau 4b), l'indexation par bigrammes ou par des mots après segmentation automatique semble donner des performances similaires d'un point de vue statistique. Si, pour le corpus chinois, les différences sont marginales, on notera tout de même un léger avantage pour l'indexation par mots pour la langue japonaise (en moyenne de 4,7 %).

Pour la langue chinoise, nous pouvons confirmer que l'indexation par des bigrammes propose une meilleure performance qu'une indexation par unigrammes (ou caractères), une différence relative en moyenne de 19 % (voir tableau 4a). Cependant, le test du signe indique habituellement que ces différences de performance ne sont pas significatives. Pour cette langue, la meilleure indexation semble être une approche combinée “unigrammes & bigrammes“ (tableau 4b). Ce choix permet, en moyenne, une augmentation de la précision moyenne de 10 % par rapport à une indexation par mots ou par bigrammes.

Pour la langue coréenne, la décomposition en morphèmes permet d'améliorer significativement la performance moyenne comparée à une indexation par mots simples (+ 48 % dans le tableau 5). Une indexation par bigrammes apporte également une précision moyenne significativement plus élevé qu'une indexation par mots simples (en moyenne de 57 %). Finalement, si l'on compare les indexations par morphèmes ou par bigrammes, nos expériences semblent indiquer que les bigrammes proposent une meilleure précision moyenne, sans que la différence entre ces deux indexations soit statistiquement significative.

Remerciements

Cette recherche a été subventionnée, en partie, par le Fonds National Suisse (subside numéro 200020-103420). L'auteur remercie J. Savoy pour ses suggestions et remarques lors de la rédaction de cet article.

Bibliographie

Amati G., van Rijsbergen C.J., « Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness », *ACM-TOIS*, vol. 20, n° 4, 2002, p. 357-389.

Chen A., Gey F.C., « Experiments on Cross-Language and Patent Retrieval at NTCIR-3 workshop », *Proceedings NTCIR-3*, Tokyo, 2003, p. 216-224.

Kishida K., Chen K.H., Lee S., Kuriyama K., Kando N., Chen H.-H., Myaeng S. H. Demeure I., Farhat J., « Overview of CLIR Task at the Fifth NTCIR Workshop », *Proceedings NTCIR-5*, Tokyo, 2005, p. 1-10.

Kwok K.L., « Employing Multiple Representations for Chinese Information Retrieval », *JASIST*, 1999, New York, John Wiley & Sons, vol. 50, n° 8, p. 709-723.

Kwok K.L., Dinstl N., Choi S., « NTCIR-4 Chinese, English, Korean Cross Language Retrieval Experiments using PIRCS », *Proceedings NTCIR-4*, Tokyo, 2004. p. 186-192.

Lee J.H., Ahn J. S., « Using n-grams for Korean Text Retrieval », *Proceedings ACM-SIGIR*, Zurich, 1996, ACM Press, p. 216-224.

Luk R.W.P., Kwok K.L., « A Comparison of Chinese Document Indexing Strategies and Retrieval Models », *ACM-TALIP*, vol. 1, n° 3, 2002, p. 225-268.

Matsumoto Y., Kitauchi A., Yamashita T., Hirano Y., Matsuda H., Asahara M., « Japanese Morphological Analysis System ChaSen », NAIST (http://chasen.aist-nara.ac.jp/).

McNamee P., « Experiments in the Retrieval of Unsegmented Japanese Text at the NTCIR-2 Workshop », *Proceedings NTCIR-2*, Tokyo, 2001, p. 157-162.

Nie J.-Y., Ren F., « Chinese Information Retrieval: Using Characters or Words? », *Information Processing & Management*, vol. 35, n° 4, 1999, p. 443-462.

Nie J.-Y., Gao J., Zhang J., Zhou M., « On the Use of Words and N-grams for Chinese Information Retrieval », *Proceedings IRAL*, Hong Kong, 2000, ACM Press, p. 141-148.

Peters C., « Working Notes for the CLEF 2005 Workshop », available at the Web site http://www.clef-campaign.org/2005/working_notes/ (Visited, December, 5th, 2005).

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.

Savoy J., « Report on CLEF-2002 Experiments: Combining Multiple Sources of Evidence », *Advances in Cross-Language Information Retrieval*, LNCS vol. 2785, Springer-Verlag, Berlin, 2003, p. 33-90.

Annexe 1. Spécification exacte des paramètres

Stratégie d'indexation	*b*	k_1	*avdl*	*c*	*slope*	*pivot*
chinois (unigramme)	0,55	2,0	384	1,0	0,1	179
chinois (bigramme)	0,55	2,0	421	1,0	0,1	321
chinois (uni. & bigram.)	0,55	2,0	805	1,0	0,1	498
chinois (mots)	0,55	2,0	267	1,0	0,1	168
japonais (bigramme)	0,4	1,2	196	6,0	0,1	133
japonais (mot)	0,6	1,2	144	6,0	0,1	92
coréen (mot)	0,55	1,2	176	3,0	0,1	143
coréen (mot HAM)	0,55	2,0	223	3,0	0,1	156
coréen (bigramme)	0,75	2,0	333	1,0	0,1	233

Tableau A1. *Paramètres des modèles Okapi, PB2, « Lnu-ltc » et « dtu-dtn »*